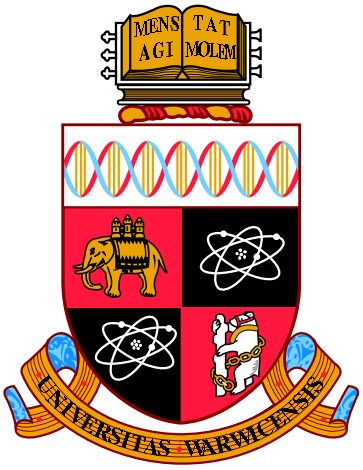


# Bayesian Hierarchical Stochastic Inference on Multiple, Single Cell, Latent States from both Longitudinal and Stationary Data

by

**Simone Tiberi**

**Thesis**

Submitted to the University of Warwick

for the degree of

**Doctor of Philosophy**

**Department of Statistics**

DECEMBER 2016

THE UNIVERSITY OF WARWICK

He's more myself than I am.
Whatever our souls are made of,
his and mine are the same.

---

*Wuthering Heights, Emily Brontë*

*A te*

# Contents

# List of Tables

# List of Figures

# Acknowledgments

There are many people who contributed to this work, some simply with a comment, others with endless hours of work. Firstly I thank my supervisor, Prof. Bärbel Finkenstädt., for bringing me into the world of research: she gave me an incredible opportunity and invested a huge amount of time (and patience!) to guide me throughout my PhD. In addition, I thank Prof. David Rand for the long time spent in countless meetings together and for the precious knowledge and advice offered me. A big thanks also goes to Dr Daniel Hebenstreit, Dr Hiroshi Momiji, Dr Kirsty Hey and Dr Silvia Calderazzo for their precious suggestions and contributions during long conversations. I also want to mention other people who gave me useful advice and explanations in the past years, I am particularly grateful to Dr Mark Fiecas, Dr Anthony Lee, Dr Krzysztof Latuszynski, Dr David Rossell, Dr Mark Walsh, Prof. Paul Thornalley, Lucy Ternent and Dr Massimo Cavallaro. I also want thank my past panel members who helped me improving this piece of work, Dr Julia Brettschneider, Dr Dario Spanó and Dr Sara Wade.

My gratitude also goes to the institutions which gave me the possibility of carrying out this PhD: the Engineering and Physical Sciences Research Council (EPSRC), which funded my PhD, the Department of Statistics and the University of Warwick; being part of it has been a real pride and honour for me.

A thought also goes to my mother, my father and Prof Maria Grazia Pittau, who first introduced me to statistics. Last but not least, a massive thanks for the people who made this PhD way more pleasant than it would have been otherwise, in particular to Pantelis, Panayiota, Habib, Elia, Thodoris and Tolo. This thesis is also dedicated to you.

# Declarations

This Thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree.

The experimental data used in Part I has been provided by Xue *et al.* (2015b), from the Warwick Systems Biology Centre, although most of the images have been self processed to obtain the raw data needed for the analysis. The experimental data in Part II has been provided by Dr Daniel Hebenstreit, from the School of Life Sciences at the University of Warwick.

# Abstract

This thesis provides inferential methods for the analysis of two biological systems, for which single cell experimental data are available.

In the first part of the thesis we focus on a hierarchical analysis on multiple, single cell, Nrf2 reporter levels in nucleus and cytoplasm, observed in human endothelial HMEC-1 *in vitro* cells (Xue *et al.*, 2015a). Nrf2 is a transcription factor that regulates the expression of several defensive genes protecting against various cellular stresses and forms of oxidation. This analysis aims to gain an insight into this essential cellular protective mechanism. We propose a reaction network based on five reactions, including a distributed delay and a non-linear term, for longitudinal measurements of the amount of Nrf2 in nucleus and cytoplasm. The diffusion approximation (DA) is used to approximate this Markov jump process with a stochastic delay differential equation (SDDE). Since this continuous process is only observed at discrete time points, a second approximation, the Euler-Maruyama approximation (EMA) of the DA, is needed to obtain an approximate likelihood for this bivariate process. Furthermore, to make use of multiple single cell data, we embed the model in a Bayesian hierarchical framework. Moreover, a measurement equation, which involves a proportionality constant and a bivariate normal error, for the nuclear and cytoplasmic measurements, is necessary to relate the original unobserved population levels, X, to the observations, Y. This introduces a hidden Markov process for X and a Bayesian analysis is performed, via a data augmentation procedure, to explore the high dimensional posterior space which includes a bivariate latent process X for every cell. We show results obtained on simulation studies, proving the validity of the methodology, and on a real data application, composed of 35 single cell fluorescent

levels under the basal condition, and of 36 under the induction by a stimulant, both observed every two minutes for 1.5-7 hours.

In the second part of the thesis we describe the analysis of a switch gene model for mRNA populations. We consider a gene that switches, with exponential waiting times, between a more active ON state and a less active OFF state, where the gene transcribes mRNA at a higher and a lower rate, respectively. We observe, via a measurement equation, the mRNA level in each cell, which is assumed to have reached a steady state. We analytically derive the stationary distribution of such a model and infer its parameters from experimental data, again via hierarchical Bayesian inference. The mRNA populations are only observed up to a proportionality constant and with a second source of white noise attributed to the measurement process. As in the previous case, we use a data augmentation procedure to explore the posterior space of the latent data. The analysis is repeated for different levels of induction by *tetracycline*, a stimulant, which results in increased gene expression. We particularly focus on studying how the stimulation affects the system.

# Abbreviations

| | |
|---|---|
| ACF | Autocorrelation |
| AR | Autoregressive |
| ARE | Antioxidant response element |
| ARW | Adaptive random walk |
| CI | Credible interval |
| CLE | Chemical Langevin equation |
| CME | Chemical master equation |
| DA | Diffusion approximation |
| DDE | Delay differential equation |
| DE | Differential equation |
| DNA | Deoxyribonucleic acid |
| EM | Euler-Maruyama |
| EMA | Euler-Maruyama approximation |
| FC | Flow cytometry |
| FISH | Fluorescence in situ hybridization |
| HBB | Hemoglobin subunit beta |

| | |
|---|---|
| HIV | Human immunodeficiency virus |
| HIV-1 | Human immunodeficiency virus type 1 |
| HMM | Hidden Markov model |
| HPD | Highest posterior density |
| iid | Independent and identically distributed |
| LB | Lower bound |
| LNA | Linear noise approximation |
| LRT | Likelihood ratio test |
| MC | Monte Carlo |
| MCCM | Markov chain Monte Carlo |
| MH | Metropolis-Hastings |
| MJP | Markov jump process |
| MLE | Maximum likelihood estimate |
| MM | Michaelis-Menten |
| mRNA | Messenger ribonucleic acid |
| NFE2L2 | Nuclear factor (erythroid-derived 2)-like 2 |
| Nrf2 | Nuclear factor (erythroid-derived 2)-like 2 |
| ODE | Ordinary differential equation |
| PACF | Partial autocorrelation |
| PMMH | particle marginal Metropolis-Hastings |
| $p$Nrf2 | Phosphorylated Nrf2 |

| | |
|---|---|
| RHS | Right hand side |
| RNA | Ribonucleic acid |
| r.v. | Random variable |
| RW | Random walk |
| SDDE | Stochastic delay differential equation |
| SDE | Stochastic differential equation |
| SFN | Sulforaphane |
| TF | Transcription factor |
| UB | Upper bound |
| WT | Wild type |

# Chapter 1

# Introduction

> Science is but a perversion of itself unless it has as its ultimate goal the betterment of humanity.
>
> *Nikola Tesla*

Recent advances in fluorescent light reporter techniques have allowed biologists to collect good quality data at a single cell level. This thesis presents original methods for inference in two stochastic biological systems, both modelling transcriptional dynamics in single cells. In Part I we consider a stochastic system for protein populations where, for each cell, we observe repeated measurements, in both nucleus and cytoplasm, generating a bivariate longitudinal process. Conversely, in Part II we infer a dynamical system for mRNA populations where the system is assumed to be in steady state and mRNA levels are observed once for each cell. In both cases, a measurement equation is used to relate the observations to the original unobserved levels, which are treated as latent states. In order to deal with these hidden layers, we use data augmentation methods to sample the unobserved data. Inference is carried out, via Metropolis-within-Gibbs algorithms, by alternately sampling from the conditional distributions of parameters and latent states. Both analyses follow a Bayesian hierarchical approach which allows us to explicitly model the experimental variability and to compare different experimental conditions by means of the hyper-parameters' posterior densities.

A significant portion of the thesis is dedicated to the analysis of the experimental data.

## 1.1 Thesis outline

### Part I

In the first Part of the thesis we consider a collection of multiple single cell bivariate processes, each describing the levels of Nrf2 protein in the nucleus and cytoplasm of a cell. Nuclear factor (erythroid-derived 2)-like 2, also known as NFE2L2 or Nrf2, is a transcription factor that regulates the expression of detoxifying and anti-oxidant enzyme genes (Itoh *et al.*, 2003), protecting against environmental toxic attacks, oxidative stress, lipid peroxidation, macromolecular damage, metabolic dysfunction and inflammation (Taguchi and M, 2011; Malhotra *et al.*, 2010). Recently, it was found (Xue *et al.*, 2015a,b) that Nrf2 undergoes autonomous translocations, independently of stimulation, between nucleus and cytoplasm, which generate oscillations in the levels of nuclear and cytoplasmic Nrf2. Our aim is to introduce a mathematical model for such oscillations, and infer its parameters from experimental data, to gain an insight into the nature of this protective mechanism.

In order to analyze experimental data, we use population dynamical systems. These are systems which model the dynamics of species as birth and death processes. However, due to the analytic intractability of the resulting Markov jump process (MJP), approximations are needed. In particular we focus on the diffusion approximation (DA) which, jointly with the Euler-Maruyama approximation (EMA), allows us to formulate and compute an approximate likelihood for fitting the model to real data. The measurement equation relating the observations to the original Nrf2 reporter levels involves a proportionality constant and a bivariate normal error. Inference is implemented by alternately sampling from the conditional distributions of the model parameters and the unobserved population levels, treated as a hidden Markov process. In order to take advantage of the availability of multiple single cell data, a Bayesian hierarchical structure is assumed at the cell level. Simulations are carried out to assess the validity of the proposed methodology, which is then used to infer parameters from the available experimental data.

Part I is organised as follows. Chapter 2 introduces the biological system, the reaction network associated with it and the available data. Chapter 3 describes the likelihood approximation, the measurement equation and the inferential methodology, involving the data augmentation procedure, the hierarchical framework and the MCMC algorithm. It also illustrates simulation studies showing the validity of our methods. Chapter 4 presents inference on the experimental data, as well as some preliminary exploratory analyses, allowing us to gain a useful insight into the Nrf2 system. Chapter 5 concludes the Nrf2 study with a stability analysis of the

system near its equilibrium in order to study the qualitative nature of the observed oscillations.

## Part II

The second part of the thesis focuses on the analysis of mRNA levels in single cells, for a large population of cells. We propose a two-state switch model for the gene expression: the gene switches between an active ON state and a dormant OFF state; it transcribes mRNA at a higher and lower level in the former and latter state, respectively. We find that such a system has a Poisson-beta distribution for the population of mRNA in a cell at equilibrium.

As in the previous analysis, we introduce a measurement equation, involving a proportionality constant and a source of random noise, to relate the observations to the unobserved populations of mRNA, which are treated as latent states. Again, as for the Nrf2 analysis, we employ a data augmentation procedure to deal with the unobserved data and we alternately sample from the conditional distributions of parameters and latent states. The experimental data consist of a single measurement in each cell, which is assumed to be at equilibrium, for a multitude of cells, about 10,000 in each experiment. Given the availability of four experimental replicates the analysis is embedded in a Bayesian hierarchical framework, with a hierarchy on the experimental replicate. Furthermore, data are collected under various levels of induction by *tetracycline*, a stimulant, which appears to boost gene expression levels. After showing the validity of our methodology in simulation studies, we use it to infer parameters from the available data and compare results under different experimental conditions.

Part II is organised as follows. Chapter 6 introduces the biological framework, the motivation of the analysis, the available data which will be used for inference and the two-state switch model we propose. Chapter 7 mostly focuses on the mathematical and statistical technical aspects: the derivation of the Poisson-beta stationary distribution for the mRNA population, the measurement equation, the data augmentation procedure, the hierarchical framework and the MCMC algorithm. In Chapter 7 we also present results from simulation studies indicating the methodology we propose is a valid inferential tool. Chapter 8 shows inference on the available data and draws some conclusions about the nature of the biological mechanism by comparing posterior results for the different experimental conditions.

### Conclusions

At the end of the thesis we briefly summarize the results and achievements obtained in both analyses, introduce ideas about possible extensions of the work presented here and conclude the thesis.

# Part I

# Modelling and Bayesian inference for single cell Nrf2 latent processes

# Chapter 2

# Biological processes

> There are many things in life that will catch your eye, but only a few will catch your heart. Pursue these.
>
> *Michael Nolan*

In this Chapter we introduce some background notions about Nrf2 and explain the motivation of the study. We then describe how the experimental data is obtained and present a simple modelling approach for Nrf2 movements; we also present the parameters of interest and the reaction network which will be used in the following Chapters, to build a likelihood for the model and analyse our data.

## 2.1 Motivation

The excess of oxidative stress is implicated in several diseases, such as cancer, diabetes, atherosclerosis, Alzheimer's and Parkinson's (Pi *et al.*, 2007). The induction of genes that protect against such stress is essential to control the oxidative damage. These genes are regulated through antioxidant response elements (AREs) by Nuclear factor (erythroid-derived 2)-like 2, simply called NFE2L2 or Nrf2. Nrf2 is a transcription factor; in other words it is a protein which regulates the expression of other genes. In particular Nrf2 plays a central role in this protective mechanism by up-regulating ARE-driven detoxification and antioxidant genes, especially if under specific stimuli (Lee *et al.*, 2005).

The target of our study is to develop an inferential procedure to model the dynamical Nrf2 movements between nucleus and cytoplasm of a cell, in order to improve our current knowledge of the nature of this essential phenomenon.

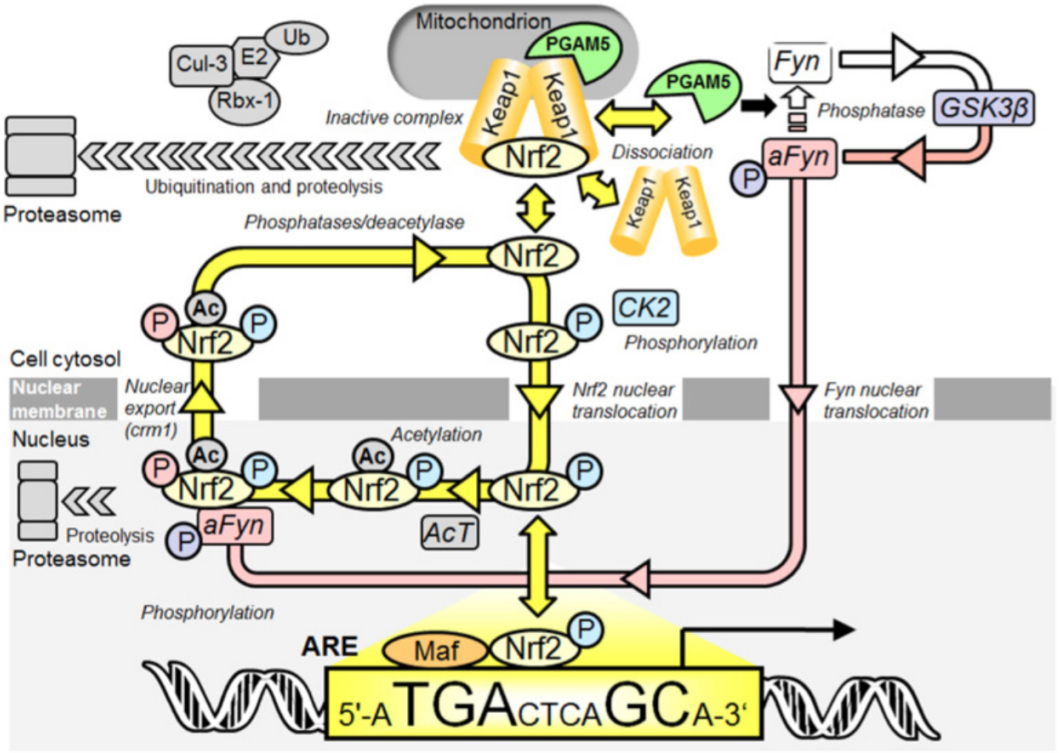


**Figure 2.1:** Biological mechanism generating Nrf2 oscillations. Reproduced from Xue *et al.* (2015b).

Figure 2.1 graphically illustrates the relevant biological processes. In the cytoplasm, Nrf2 regulatory activity is repressed by binding to the proteins Keap1 and PGAM5, which act as inhibitors, holding Nrf2 in the cytoplasm (Kobayashi *et al.*, 2004; Itoh *et al.*, 2003). When the PGAM5:Nrf2:Keap1 complex is disrupted, the liberated Nrf2 enters the nucleus and activates its target genes by binding to AREs (Pi et al., 2007). This process is enhanced when the cell is exposed to oxidative and electrophilic stress. At the same time, the release of PGAM5 from the complex decreases phosphatase activity (Lo and Hannink, 2006; Takeda *et al.*, 2009): PGAM5 inhibits Fyn de-phosphorylation and thus effectively increases the active phosphorylated Fyn, that we call *a*Fyn, which enters the nucleus. Thereafter Nrf2 is phosphorylated by the *a*Fyn kinase and expelled from the nucleus to the cytoplasm (Kaspar and Jaiswal, 2010), where it is bound once again by Keap1 and PGAM5.

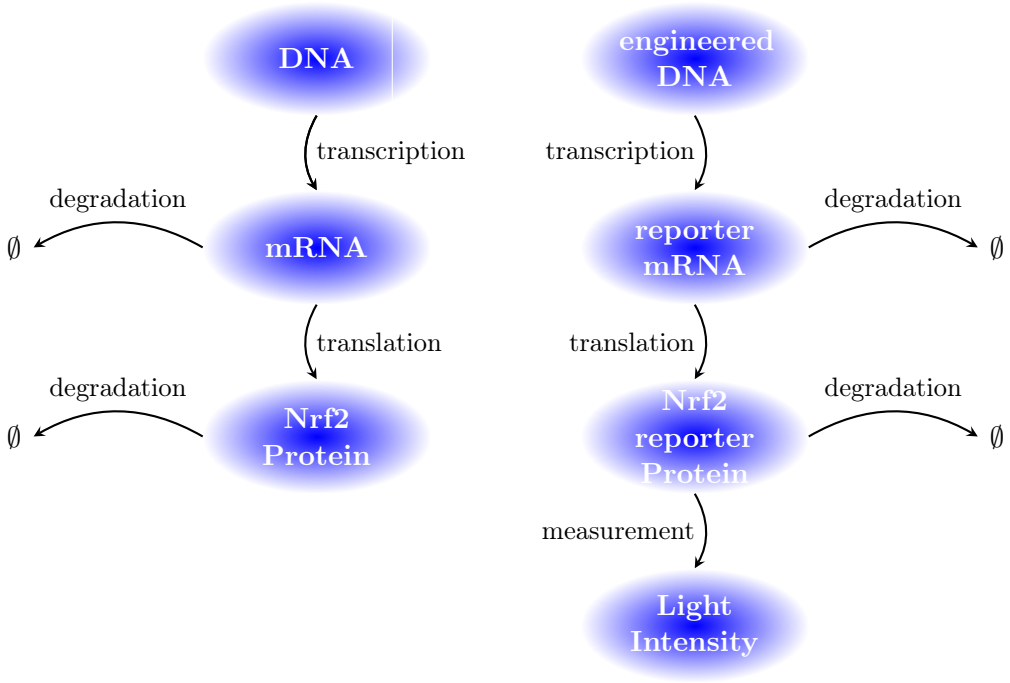


**Figure 2.2:** Measurement process for the Nrf2 reporter protein.

## 2.2 Data

The available data, acquired from Xue *et al.* (2015b) although most images are self processed to obtain the raw data, have been obtained from fluorescent images of a reporter of Nrf2 protein in human endothelial HMEC-1 *in vitro* cells, recorded every 2 minutes for a period of about 1.5-7 hours. The measurement process, which leads to the observed light intensity, can be seen in Figure 2.2 and consists of a transient transfection of exogenous Nrf2 DNA fragments which produce an engineered reporter protein able to emit light (Xue *et al.*, 2015b). First, some genetically manipulated fragments of DNA are inserted in the nucleus of the cell: from these modified filaments, Nrf2 reporter mRNA is transcribed; this mRNA translates to reporter protein, able to emit light when stimulated with some laser. Therefore the manipulation generating the reporter happens at the bottom of the process and what is actually measured is not the original protein population level (on the left of the image). Although we are interested in the latter process, it is sensible to analyse the former one because we believe reporter Nrf2 to behave in a very similar fashion to the original Nrf2 protein; hence, we assume we can obtain information about the mechanism of the Nrf2 system by analysing its reporter protein.

The light intensity obtained via this procedure is then evaluated, separately

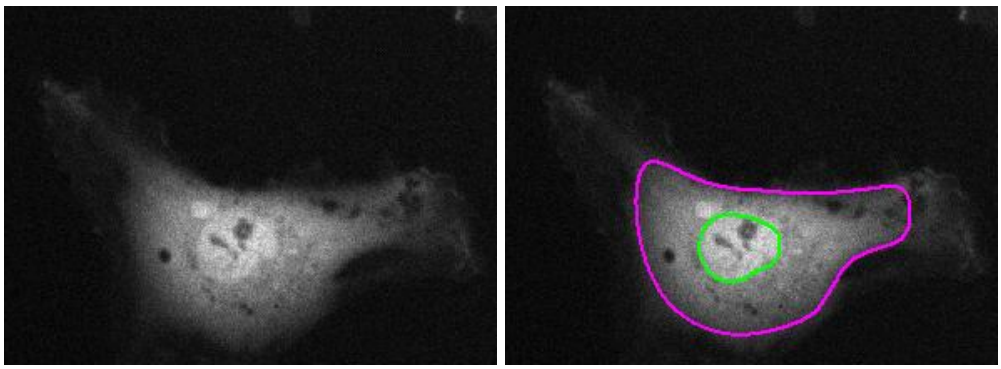

**Figure 2.3:** Light intensities of Nrf2 reporter in a cell before (left) and after (right) identifying the boundary between nucleus and cytoplasm (in green) and the cell limits (in purple).

in nucleus and cytoplasm, and converted via the *CellTracker* software (Du and Bretschneider, 2013), up to a proportionality constant and a source of white noise, into the concentration of Nrf2 reporter in each compartment. The nuclear and cellular borders need to be traced via a semi-manual procedure. As visible in Figure 2.3, one can distinguish nucleus and cytoplasm (left image) by their different light intensities and trace the border between the two compartments (green light on the right image). In a similar fashion it is possible to see the border of the cell as the light intensity rapidly disappears. This procedure is repeated every 10 frames (corresponding to 20 minutes). The *CellTracker* software (Du and Bretschneider, 2013) then interpolates the borders, also accounting for the cell movement, for the 9 frames in between. The borders of each frame obtained from this procedure are then checked by eye, one by one, and adjusted in order to minimize the measurement error. It is intuitive that such a measurement process is affected by a source of error, which we will treat with different modelling assumptions, also investigating what impact they have on inference.

Figure 2.4 shows an example of the observed dynamic levels of Nfr2 reporter in nucleus and cytoplasm, for two different cells. Interest lies in modelling the observed oscillations in the concentration of Nrf2 reporter, generated by the translocations it undergoes between nucleus and cytoplasm.

Data are collected under two experimental settings: in a basal condition and under the induction by a stimulant, *sulforaphane* (SFN). SFN is an exogenous activator (Dinkova-Kostova *et al.*, 2002; Tanigawa *et al.*, 2007) found in bioactive compounds in fruit and vegetables and the level of induction used (2 $\mu$M) is compatible with those found naturally in these aliments. For the basal and stimulated conditions, data from 35 and 36 single cells are available, respectively. Our infer-

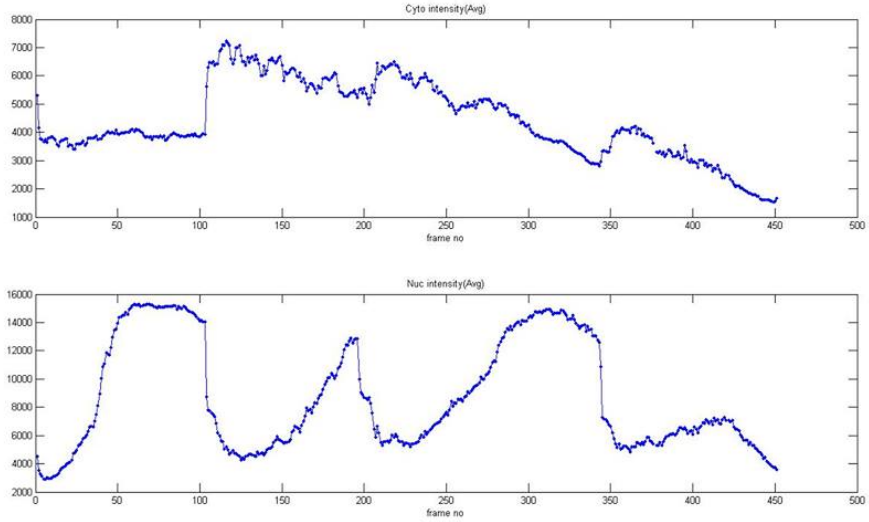


**Figure 2.4:** Observed Nrf2 oscillations in nucleus (bottom) and cytoplasm (top) of a cell.

ential method is applied to both conditions, and results under different settings are compared.

The number of images of each cell varies between 111 and 417, for the basal condition, and between 102 and 361, for the stimulated one. The average number of observations per cell are 222 and 193, again for the basal and stimulated conditions, respectively. In the Appendix, Figures B.1-B.10 show the Nrf2 concentrations observed in the available data.

Our analysis focuses on Nrf2 reporter protein levels. However, for simplicity, in the following we will mostly refer to it as Nrf2.

## 2.3 Biochemical modelling

We now introduce the notion of a biochemical reaction network, a concept that will allow us to appropriately define our system. A biochemical reaction network network is a system involving $r$ species in a well stirred environment, $X^{(1)}, ..., X^{(r)}$, taking values in $\mathbb{N}$, and $q$ reactions $R_1, ..., R_q$, where each reaction $R_i$ is defined as

$$R_i : m_{i,1}X^{(1)} + m_{i,2}X^{(2)} + ... + m_{i,r}X^{(r)} \rightarrow n_{i,1}X^{(1)} + n_{i,2}X^{(2)} + ... + n_{i,r}X^{(r)} \quad (2.1)$$

or, analogously, in vector form,

$$R_i : m_i X \to n_i X, \tag{2.2}$$

with $X = (X^{(1)}, ..., X^{(r)})^T$ and where $m_i = (m_{i,1}, ..., m_{i,r})$ and $n_i = (n_{i,1}, ..., n_{i,r})$ respectively indicate the number of molecules of each species involved as reactants and products of the $i$-th reaction (Wilkinson, 2012). Normally, $m_i$ and $n_i$ assume integer values, as it is the case in here, and are called stoichiometries.

We define the stoichiometry vector associated to $R_i$, as $v_i = n_i - m_i, i = 1, ..., q$. This vector indicates the changes generated by the corresponding reaction in the populations of species. From these, we can introduce the $r \times q$ stoichiometry matrix $S = (v_1^T, ..., v_q^T)$.

Every reaction $R_i$ is associated to a hazard $w_i(X_t, \theta)$ which depends on the value of the process at time t, $X_t = (X_t^{(1)}, ..., X_t^{(r)})^T$, and on a vector of constants, $\theta$, which represents our kinetic parameter vector. Each hazard indicates how likely the corresponding reaction is to happen in an infinitesimal time interval. Considering the infinitesimal time $dt$, then $w_i(X_t, \theta)\, dt$ indicates the probability that reaction $R_i$ occurs in $(t, t + dt]$, $i = 1, ..., q$ (Wilkinson, 2012).

## 2.4 Nrf2 reaction network

The biological process described in Section 2.3 naturally generates translocations between nucleus and cytoplasm which can produce oscillations in nuclear and cytoplasmic levels of Nrf2. By looking at Figures B.1 to B.10 in the Appendix, we can see that several cells, although not most of them, exhibit clear oscillations in their nuclear and cytoplasmic concentrations. We now propose a two dimensional system which is able to explain the observed oscillations and takes into account the intracellular movements of Nrf2 proteins, between nucleus and cytoplasm, as well as their synthesis and degradation. The model is a simplified representation of the structure described before: in particular, it simplifies Nrf2 nuclear phosphorylation, acting via the effect of *a*Fyn, with a delay term. Several delay differential equation (DDE) models are proposed by (Xue *et al.*, 2015b). In collaboration with some of the authors of the paper (Xue *et al.*, 2015b), Dr Hiroshi Momiji and Prof. David Rand (both at Warwick Systems Biology Centre) and Prof. Paul Thornalley (Warwick Medical School), we modify those DDEs and develop a model which simplifies the system, facilitating parameter inference while reproducing its main feature, i.e. oscillations. In particular, the model we present here represents the final compromise between the practical identifiability of parameters, observed in simulations, and

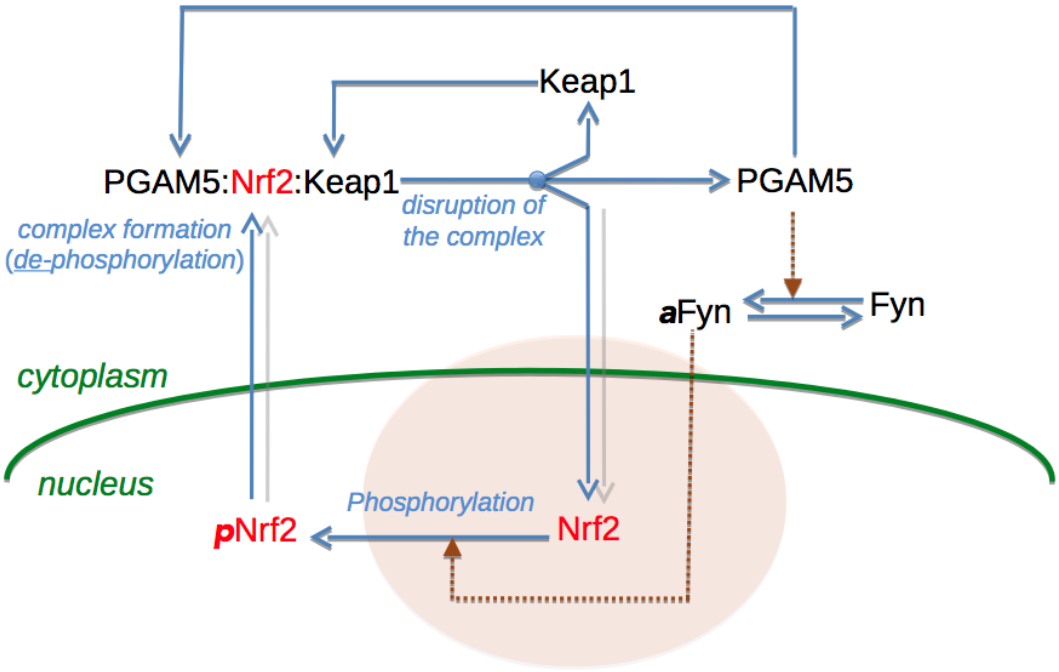


**Figure 2.5:** Simplified version of biological mechanism driving Nrf2 translocations: the blue solid lines indicate transformations, the red dashed lines represent signaling. The phosphorylated nuclear Nrf2, $p$Nrf2, is substituted by the delayed nuclear level.

the model realism. Indeed, every modification aimed at easing inference, has been discusses with the collaborators cited above before being implemented. The system we use in our analysis is graphically illustrated in Figure 2.5.

The formulation we present, illustrated in Table 2.1, involves five reactions ($q = 5$) and two populations ($r = 2$), with $X = (X^N, X^C)^T$, representing the total molecular populations of Nrf2 reporter protein in nucleus and cytoplasm, respectively, and their value at time $t$ is denoted by $X_t = (X_t^N, X_t^C)^T$.

We now illustrate the reactions and respective rates and stoichiometry vectors. All parameters used in the hazards are real positive rates, which we embed in a parameter vector that we call $\theta$, that will be defined later, once all parameters have been introduced. All rates are expressed per minute.

- $R_1$ represents the movement of a molecule of Nrf2 reporter from the cytoplasm to the nucleus, which generates an increase of 1 in $X^N$ and a decrease of the same amount in $X^C$, hence $v_1 = (1, -1)$. The respective rate is linear and proportional to the population of cytoplasmic Nrf2: $w_1(X_t, \theta) = k_d\, X_t^C$.

- $R_2$ indicates the movement of a molecule in the opposite direction, i.e. from the

nucleus to the cytoplasm, generating an increase in the cytoplasmic amount and a decrease in the nuclear one, thus $v_2 = (-1, 1)$. This reaction has a non-linear hazard function. This has a key role in that it is responsible for the oscillations in the nuclear and cytoplasmic levels, also under the deterministic setting, as we will show in Section 2.5. In particular, we assume that the rate for $R_2$ involves a Michaelis-Menten (MM) formulation, which represents a special case of the Hill equation where the Hill coefficient has been set to one. The MM term is $\frac{k_a\, X_t^N}{K_a + X_t^N}$, where, $k_a$ represents the maximum rate achieved by the system, attained at the substrate saturation concentration, and $K_a$, the Michaelis constant, is equal to the concentration of substrate at which the reaction rate is half the maximum, $k_a$ (Down and Riggs, 1964; Goldbeter, 1996). The rate for $R_2$ further involves a distributed delay $\tau$, representing the time from the moment Nrf2 enters the nucleus to when it is phosphorylated. The delay term represents an integral over the distribution of the delay $f_\tau(.)$, which will be specified later, $E_\tau(X_{t-\tau}^N) = \int X_{t-\tau}^N\, f_\tau(\tau) d\tau$, where $E_\tau(.)$ denotes the expectation operator with respect to the distribution of $\tau$ and $\tau^{max}$ is the maximum value the delay is assumed to take.

We define the rate for reaction $R_2$ as:

$$w_2(X_t, ..., X_{t-\tau^{max}}, \theta) = E_\tau(X_{t-\tau}^N)\, \frac{k_a\, X_t^N}{K_a + X_t^N},$$

where, to simplify the notation, we introduce $X_{t,...,t-\tau^{max}} = (X_t, ..., X_{t-\tau^{max}})$ as the element containing all the delayed populations up to the maximum delay, from time $t$ to $t - \tau^{max}$.

The term $E_\tau(X_{t-\tau}^N)$ models the delayed nuclear export signal (NES), originating in the reversible phosphorylation of Fyn, and approximates the fraction of phosphorylated nuclear Nrf2, denoted by the symbol $p$Nrf2 in Figure 2.5. In fact, $p$Nrf2 is a subset of the total nuclear Nrf2 and hence can be approximated by the delayed nuclear Nrf2 multiplied by a factor $\tilde{c}$, $0 < \tilde{c} < 1$, representing the fraction of $p$Nrf2 over the total nuclear Nrf2, which is assumed to be constant over time. Therefore, when interpreting inference about $k_a$, one should keep in mind that this parameter represents the product of the original Michaelis-Menten parameter, that we call $\bar{k}_a$, and the constant $\tilde{c}$: $k_a = \bar{k}_a\, \tilde{c}$.

Similar to $R_2$, we initially considered a non-linear rate for $R_1$ too. However, we noticed in simulations that such a system, with two non-linear rates, leads to difficulties in identifying the corresponding highly correlated parameters.

Therefore, together with our collaborators, we decided to simplify $w_1$ with a linear rate; instead, we kept $w_2$ in its non-linear form due to the greater biological complexity of the nuclear export, which is believed *a priori* to drive the observed oscillations and would not be realistically modelled by a linear rate.

- $R_3$ denotes the synthesis of one new protein in the cytoplasm, causing an increment of 1 in $X^C$, so $v_3 = (0, 1)$. In principle, the synthesis of new protein depends on the amount of Nrf2 reporter mRNA. However, since we do not observe these mRNA levels, we approximate the associated hazard, with a constant rate: $w_3(X_t, \theta) = \gamma$.

- $R_4$ and $R_5$ refer to two different degradation pathways, happening in nucleus and cytoplasm. They both generate a loss of 1 molecule in the respective compartment, hence $v_4 = (-1, 0)$ and $v_5 = (0, -1)$. The rates, governed by the same constant $\delta$, representing the cellular degradation rate per molecule, are proportional to the molecular populations of Nrf2 in either compartment and are equal to $w_4(X_t, \theta) = \delta\, X_t^N$ and $w_5(X_t, \theta) = \delta\, X_t^C$. We also considered a more general formulation for the degradation pathway by assuming two distinct degradation rates in nucleus and cytoplasm. Such a structure introduced a strong correlation between the two degradation rates which were difficult to accurately infer. In collaboration with Prof. Paul Thornalley, we simplified the system keeping its model realism by assuming a common degradation rate in the two cellular compartments.

**Distributed delay**

We introduce parameters $\mu_\tau$ and $\sigma_\tau$, both in $[0, \tau^{max})$, representing the mean and standard deviation of $\tau$, respectively. We assume that the distribution of the delay can be realistically modelled by a scaled beta distribution with values between 0 and a fixed maximum for the delay $\tau^{max}$, $\frac{\tau}{\tau^{max}} \sim Beta\left(\frac{\mu_\tau}{\tau^{max}}, \frac{\sigma_\tau}{\tau^{max}}\right)$, where $Beta(\mu, \sigma)$ denotes the beta distribution with mean $\mu \in [0, 1]$ and standard deviation $\sigma \in [0, 1)$. Therefore we can now define $\theta = (k_d, k_a, K_a, \mu_\tau, \sigma_\tau, \gamma, \delta)$ as the kinetic parameter vector of the system, where each element belongs to $\mathbb{R}^+$.

The delayed component is obtained, at each time point $t$, as the average over the possible values of the distribution of $\tau$,

$$E_\tau(X_{t-\tau}^N) = \int_0^{\tau^{max}} X_{t-\tau}^N f_\tau(\tau) d\tau, \tag{2.3}$$

| Reaction | $R_1$ | $R_2$ | $R_3$ | $R_4$ | $R_5$ |
|---|---|---|---|---|---|
| Reaction description | Nuclear import | Nuclear export | Synthesis | Nuc. deg. | Cyto. deg. |
| Change | $X_t^C \to X_t^N$ | $X_t^N \to X_t^C$ | $\emptyset \to X_t^C$ | $X_t^N \to \emptyset$ | $X_t^C \to \emptyset$ |
| $v_i$ | (1, -1) | (-1, 1) | (0, 1) | (-1, 0) | (0, -1) |
| $w_i$ | $k_d\, X_t^C$ | $E_\tau(X_{t-\tau}^N)\, \dfrac{k_a\, X_t^N}{K_a + X_t^N}$ | $\gamma$ | $\delta\, X_t^N$ | $\delta\, X_t^C$ |

**Table 2.1:** Overview of reactions included in the reaction network and respective hazards, where for simplicity of notation we use $w_i$ to refer to $w_i(X_{t,\dots,t-\tau^{max}}, \theta)$. "Nuc. deg." and "Cyto. deg." are abbreviations for nuclear degradation and cytoplasmic degradation, respectively.

where the density of the delay, $f_\tau(\tau) \overset{d}{=} f_{Beta(\frac{\mu_\tau}{\tau^{max}}, \frac{\sigma_\tau}{\tau^{max}})}\left(\dfrac{\tau}{\tau^{max}}\right)/\tau^{max}$, is assumed to follow a beta density re-scaled with respect to $\tau^{max}$, where $f_{Beta(\frac{\mu_\tau}{\tau^{max}}, \frac{\sigma_\tau}{\tau^{max}})}(.)$ indicates the density of a beta distribution with mean $\mu_\tau$ and variance $\sigma_\tau$. The more usual parametrization of the beta distribution, in terms of parameters which are often called $\alpha$ and $\beta$, can easily be retrieved.

This upper limit is justified by the fact that $\tau$ represents the time from the moment Nrf2 enters the nucleus to the moment it is phosphorylated, which is only a part of the entire cycle in Figure 2.5, i.e. looking at the experimental data (Figures B.1 to B.10 in the Appendix), when a cell exhibit a clear oscillation, it hardly lasts for much longer than 1 hour; therefore we initially set $\tau^{max} = 60$ minutes. Inference on experimental data, in Section 4.6, confirms the plausibility of this choice as posterior densities for $\mu_\tau$ are far from $\tau^{max}$. Therefore it seems sensible not to extend the delay further, also considering that the first $\tau^{max}$ minutes of the observed process cannot be used in the analysis as their delayed process is not yet observed; e.g. the delay for time $t$ would be between time $t$ and time $t - \tau^{max}$, which is only available for $t \geqslant \tau^{max}$. Assuming $\tau^{max} = 60$, given data are available every 2 minutes, we can only compute the delay from the 31st observations.

The scaled beta is a flexible distribution that naturally constrains the delay to its support $[0, \tau^{max}]$, without the need for any truncation, that would be required if we used the gamma distribution, as in Heron *et al.* (2007).

Since the process is only observed at discrete time points, the integral in (2.3)

is approximated by the sum over small intervals of the observed process weighted by the density of the delay:

$$E_\tau(X^N_{t-\tau}) = \sum_{\tau=0}^{\tau^{max}} X^N_{t-\tau}\, w_\tau(\tau), \text{ where } w_\tau(\tau) = \frac{f_\tau(\tau)}{\sum_{\tau'=0}^{\tau^{max}} f_\tau(\tau')}. \tag{2.4}$$

The denominator of the weights $w_\tau(\tau)$ is needed to rescale for the fact that the sum of the densities, at discrete time points, does not add to one.

## 2.5 DDE study

Our preliminary Ansatz is that the observed oscillations in Nrf2 levels are due to the intrinsic nature of the biological mechanism which generates Nrf2 intracellular movements. In other words, we believe these oscillations are not simply due to the stochasticity of the process, arising from a limited number of molecules, and that we would observe them even in an, ideally, infinite population of molecules (Xue *et al.*, 2015b). Therefore we investigate the behaviour of the system under the deterministic setting, which can be interpreted as the mean trend of the process, in order to study the model when the population tends to infinity.

The deterministic model associated to the system in Table 2.1 is defined by the following delay differential equation (DDE),

$$\dot{Z}^N_t = k_d\, Z^C_t - E_\tau(Z^N_{t-\tau})\, \frac{k_a\, Z^N_t}{K_a + Z^N_t} - \delta Z^N_t \qquad [\text{ Nuc Nrf2 }] \tag{2.5}$$

$$\dot{Z}^C_t = E_\tau(Z^N_{t-\tau})\, \frac{k_a\, Z^N_t}{K_a + Z^N_t} - k_d\, Z^C_t + \gamma - \delta Z^C_t, \qquad [\text{ Cyt Nrf2 }] \tag{2.6}$$

where $Z^N_t$ and $Z^C_t$ represent the deterministic counterpart of $X^N_t$ and $X^C_t$ respectively, the means of the stochastic process for the evolution of the molecular population, while $E_\tau(Z^N_{t-\tau})$ is the integral in (2.3) applied to $Z^N$. The delay component, $E_\tau(Z^N_{t-\tau})$, which is an integral over the delay distribution, is approximated by a discrete summation over the past values of the process, up to lag $\tau^{max}$, as in (2.4).

We solve the DDE in  (R Core Team, 2016), via a numerical solver (Soetaert *et al.*, 2016), and show that, under some parameter values, the DDEs in (2.5) and (2.6) are able to generate, as required, sustained oscillations in the nuclear and cytoplasmic populations (see Figure 2.6). This indicates that our model is able to explain an oscillatory behaviour which is intrinsic in the nature of the biological system and, hence, reproduce it even in a deterministic setting.

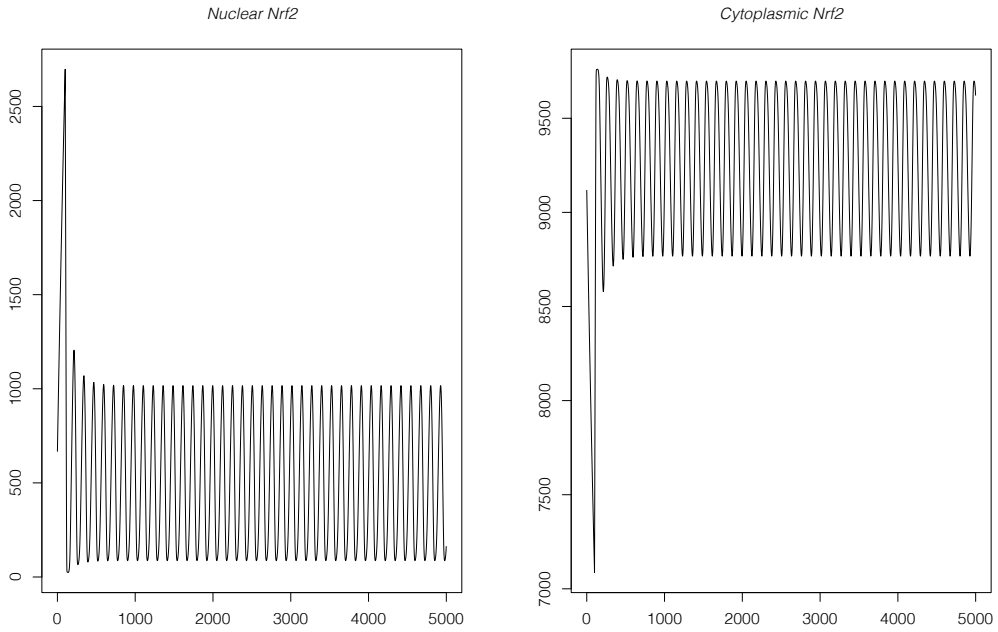


**Figure 2.6:** DDE study of our reaction network generating sustained oscillations in time, under a specific set of parameter values. Parameter values used to reproduce the image in the picture are as follows: $k_d = 3.04\,10^{-3}, k_a = 9.25\,10^{-2}, K_a = 166.52, \mu_\tau = 25.85, \sigma_\tau = 3, \gamma = 1.07\,10^{-1}$ and $\delta = 1.10\,10^{-5}$.

## 2.6 Summary

In this Chapter we have introduced the biological background and the motivation of the study about the Nrf2 system. We have described how the fluorescent reported data are obtained, via a transient transfection of engineered DNA fragments. We have also presented a reaction network model for the population of Nrf2 protein in nucleus and cytoplasm in a single cell and shown that such a system can generate sustained oscillations even in a deterministic setting.

In the next Chapter we will describe the likelihood associated with this system and illustrate the inferential methodology which will then be applied to the available experimental data.

# Chapter 3

# Statistical modelling

> Pour us your poison to comfort us!
> We want, as the fire burns our
> brains, to plunge into the abyss,
> Hell or Heaven, does it matter?
> At the bottom of the Unknown to
> find something new!
>
> *Charles Baudelaire*

In this Chapter we introduce an approximate likelihood for our model. We also present a measurement equation, relating the observations to the original protein levels, and embed the model in a Bayesian hierarchical framework, to take advantage of the availability of multiple single cell data. Furthermore, we describe the Metropolis-within-Gibbs algorithm we use for inference. Finally we validate our methodology via simulation studies, which show that the method is able to recover the original parameter values.

## 3.1 Diffusion approximation

A biochemical reaction network system, defined in Section 2.3, can be formulated as a Markov Jump Process (MJP) (Wilkinson, 2012; Stathopoulos and Girolami, 2013), and by considering an infinitesimal time $dt$, the probability that reaction $R_i$ occurs in the interval $(t, t + dt]$ is given by $w_i(X_{t,\ldots,t-\tau^{max}}, \theta)dt$.

However, the continuous time MJP is only observed at discrete time points; moreover, the hazards change over time due to their dependence on the value of the process at time $t$, $X_t = (X_t^N, X_t^C)^T$. As a consequence, the chemical master equation (CME) describing the evolution of the transition densities is typically analytically

intractable (Stathopoulos and Girolami, 2013; Finkenstädt *et al.*, 2013). Here we consider an approximation to the likelihood, i.e. the diffusion approximation (DA); nonetheless this does not fully overcome the issue as the DA leads to analytically intractable transitions densities too. Therefore, we further approximate the DA with the Euler-Maruyama approximation which allows us to formulate an approximate likelihood for the model.

The diffusion approximation, also known as the chemical Langevin equation (CLE) (Gillespie, 1991, 2000), models the MJP as a birth and death Poisson process. The procedure we illustrate below is normally presented for a non-delay system. However, its validity also extends to systems such as ours with a finite delay (Tian *et al.*, 2007; Hasegawa, 2004; Heron *et al.*, 2007); therefore the DA is here represented as a stochastic delay differential equation (SDDE) instead of the usual stochastic differential equation (SDE).

Consider the infinitesimal time interval $(t, t+dt]$, over which the reaction rates $w(X_{t,\dots,t-\tau^{max}}, \theta) = (w_1(X_t, \theta), w_2(X_{t,\dots,t-\tau^{max}}, \theta), ..., w_5(X_t, \theta))$ are almost surely constant. The number of reactions $R_i$ occurring in that interval, which we define as $N_i^{dt} \in \mathbb{N}$, can be modeled, for each $i = 1, ..., 5$, by an independent Poisson process, $N_i^{dt} \sim Pois(w_i(X_{t,\dots,t-\tau^{max}}, \theta)dt))$ (Wilkinson, 2012).

Hence, the instantaneous change in the population at time t, $X_t$, is given by the product between the 5 dimensional random vector, $N^{dt} = (N_1^{dt}, ..., N_5^{dt})$, and the stoichiometry matrix $S$, representing the effect that the occurrence of each reaction generates in the population,

$$X_{t+dt} - X_t = S\, {N^{dt}}^T. \tag{3.1}$$

This process has mean

$$\mu(X_{t,\dots,t-\tau^{max}}, \theta) = S\, w(X_{t,\dots,t-\tau^{max}}, \theta)^T,$$

and variance

$$\Sigma(X_{t,\dots,t-\tau^{max}}, \theta) = S\, W(X_{t,\dots,t-\tau^{max}}, \theta) S^T,$$

where $W(X_{t,\dots,t-\tau^{max}}, \theta) = diag(w_1(X_t, \theta), w_2(X_{t,\dots,t-\tau^{max}}, \theta), ..., w_5(X_t, \theta))$ is the diagonal matrix of the hazards.

The following stochastic delay differential equation (SDDE) is associated to this Poisson process (Wilkinson, 2012):

$$dX_t = \mu(X_{t,\dots,t-\tau^{max}}, \theta)\, dt + \Psi(X_{t,\dots,t-\tau^{max}}, \theta)\, dW_t, \tag{3.2}$$

where $\mu(X_{t,...,t-\tau^{max}}, \theta)$ represents the infinitesimal mean vector, $\Sigma(X_{t,...,t-\tau^{max}}, \theta) = \Psi(X_{t,...,t-\tau^{max}}, \theta)\Psi(X_{t,...,t-\tau^{max}}, \theta)^T$ is the infinitesimal variance covariance matrix and $W_t$ is a 2-dimensional Brownian motion, or Wiener process, which is simply a vector of 2 independent univariate Brownian motions.

In our system, the population process is two dimensional, $X_t = (X_t^N, X_t^C)^T$ and, by considering reactions $R_1, ..., R_5$ and corresponding hazards, $w_1(X_t, \theta), ..., w_5(X_t, \theta)$, introduced in Section 2.4, we can define the associated Poisson processes $N_1^{dt} \sim Po(w_1(X_t, \theta)\, dt), ..., N_5^{dt} \sim Po(w_5(X_t, \theta)\, dt)$. We can rewrite the expression in (3.1) as:

$$X_{t+dt} - X_t = \begin{pmatrix} N_1^{dt} - N_2^{dt} - N_4^{dt} \\ N_2^{dt} - N_1^{dt} + N_3^{dt} - N_5^{dt} \end{pmatrix}, \tag{3.3}$$

which has mean

$$\mu_t\, dt = \begin{pmatrix} \left(w_t^{(1)} - w_t^{(2)} - w_t^{(4)}\right) dt \\ \left(w_t^{(2)} - w_t^{(1)} + w_t^{(3)} - w_t^{(5)}\right) dt \end{pmatrix}, \tag{3.4}$$

and variance

$$\Psi_t\, \Psi_t^T\, dt = \Sigma_t\, dt = \begin{pmatrix} \left(w_t^{(1)} + w_t^{(2)} + w_t^{(4)}\right) dt & -\left(w_t^{(1)} + w_t^{(2)}\right) dt \\ -\left(w_t^{(1)} + w_t^{(2)}\right) dt & \left(w_t^{(1)} + w_t^{(2)} + w_t^{(3)} + w_t^{(5)}\right) dt \end{pmatrix}. \tag{3.5}$$

Here, for the sake of brevity, we have used $w_t^{(1)}, ..., w_t^{(5)}$, $\mu_t$, $\Psi_t$ and $\Sigma_t$ instead of $w_1(X_t, \theta), ..., w_5(X_t, \theta)$, $\mu(X_{t,...,t-\tau^{max}}, \theta)$, $\Psi(X_{t,...,t-\tau^{max}}, \theta)$ and $\Sigma(X_{t,...,t-\tau^{max}}, \theta)$.

Substituting expressions (3.4) and (3.5) in the SDE in (3.2), we arrive at the DA (or CLE) formulation of our system.

Again the solutions of the latter are intractable, and thus a further approximation is to be considered. If we take the increments of such process in a small time interval $\Delta t$, we can interpret equation (3.2) as the limit, for $\Delta t \to 0$, of

$$\Delta X_t = X_{t+\Delta} - X_t \doteq \mu(X_{t,...,t-\tau^{max}}, \theta)\, \Delta t + \Psi(X_{t,...,t-\tau^{max}}, \theta)\, \Delta W_t, \tag{3.6}$$

which, for finite $\Delta t$ is known as the *Euler-Maruyama approximation* (Wilkinson, 2012).

By definition of 2-dimensional Brownian motion, $W_t \sim \mathcal{N}_2(0, t\, I_2)$ and $\Delta W_t \sim \mathcal{N}_2(0, \Delta t\, I_2)$, with $I_2$ representing the identity matrix of size 2 and $\mathcal{N}_2$ the bivariate normal distribution (Wilkinson, 2012). Hence,

$$\Delta\, X_t = X_{t+\Delta} - X_t \dot{\sim} \mathcal{N}_2(\mu(X_{t,...,t-\tau^{max}}, \theta)\, \Delta t,\ \Sigma(X_{t,...,t-\tau^{max}}, \theta)\, \Delta t). \tag{3.7}$$

For the sake of simplicity we will use $\Delta$ instead of $\Delta t$ from now on.

Note that, since reactions $R_1$ and $R_2$ generate, at the same time, changes in both $X_t^N$ and $X_t^C$, $\Sigma(X_{t,...,t-\tau^{max}}, \theta)$ is non-diagonal and increments $\Delta X_t$ are jointly modeled as a negatively correlated bivariate normal.

From this formulation it is possible to obtain an approximate likelihood of the observed process:

$$L_{DA}(\theta; x) = \prod_{t=1}^{T-1} \Phi(x_{t+1} - x_t | \Delta\mu(x_{t,...,t-\tau^{max}}, \theta), \Delta\Sigma(x_{t,...,t-\tau^{max}}, \theta)), \tag{3.8}$$

where $\Phi(.|\mu, \Sigma)$ represents the normal density, with mean $\mu$ and covariance matrix $\Sigma$, and $x = (x_1, ..., x_T)$ indicates the vector of realizations from the random vector $X = (X_1, ..., X_T)$, with $x_{t,...,t-\tau^{max}} = (x_t, ..., x_{t,...,t-\tau^{max}})$ for $X_{t,...,t-\tau^{max}}$. To simplify notation, we express time $t$ in units of $\Delta$ minutes, e.g. the time between two consecutive realizations $x_{t+1}$ and $x_t$ is $\Delta = 2$ minutes.

The Euler-Maruyama approximation, necessary to obtain the approximate likelihood in (3.8), relies on the assumption that reaction rates are approximately constant over the time interval $(t, t + \Delta]$, which only holds for small enough $\Delta$. We assume that time resolution of the available experimental data, which is only 2 minutes, can be considered to be good enough for the EMA to be accurate. In fact Nrf2 reporter levels mostly undergo minor changes between consecutive observations. Therefore, the reaction rates, which also depend upon them, only vary mildly over 2 minute time intervals. This avoids the need for computationally expensive methods, such as the data augmentation technique known as "bridging" (Elerian *et al.*, 2001; Durham and Gallant, 2002), in order to fulfil the EMA constant hazards assumption.

## 3.2 Concentration level and proportionality constant

The formulation used so far, as expressed in Table 2.1, refers to the microscopic scale of the model, that is, its representation at the molecular population level. However, it is often not possible to know the total number of molecules in a cell. In fact our available data refer to the concentrations of nuclear and cytoplasmic Nrf2, as is typically the case, and not to their molecular populations.

We define $\frac{X_t}{\Omega}$ as the concentration, at time $t$, of a population of $X_t$ molecules in a container with volume $V$, constant with respect to time, where $\Omega = V\, n_A$, that is, the container volume multiplied by the Avogado number $n_A = 6.023 \times 10^{23}$ (Wilkinson, 2012). Since the volume of the container is generally unknown, and our case represents no exception, it is not possible to convert from concentrations to

total number of molecules.

In our model, the nuclear and cytoplasmic concentrations refer to different compartments of the cell. Therefore we need to introduce two separate volumes: a nuclear and a cytoplasmic one, denoted by $V^N$ and $V^C$, respectively; from which we can derive $\Omega^N = V^N\, n_A$ and $\Omega^C = V^C\, n_A$. These allow us to define the nuclear and cytoplasmic concentrations, at time $t$, as $\dfrac{X_t^N}{\Omega^N}$ and $\dfrac{X_t^C}{\Omega^C}$.

In the data collection process, we measure the light intensity of Nrf2 reporter protein. However, the measurement is thought to be proportional to the actual concentration. We introduce the proportionality constant, $\tilde{\kappa}$, which is assumed not to vary between nucleus and cytoplasm since, in both, we measure the same protein.

We can merge proportionality and concentration constants in one diagonal matrix, and introduce $\tilde{X}_t^N = \dfrac{\tilde{\kappa}}{\Omega_N} X_t^N$ and $\tilde{X}_t^C = \dfrac{\tilde{\kappa}}{\Omega_C} X_t^C$. Analogously, in matrix form, we define $\tilde{X}_t = \kappa\, X_t$, where $\kappa = diag(\kappa_N, \kappa_C)$ is the diagonal matrix with diagonal elements $\kappa_N = \dfrac{\tilde{\kappa}}{\Omega_N}$ and $\kappa_C = \dfrac{\tilde{\kappa}}{\Omega_C}$, including the proportionality constant as well as the volumes.

Parameter $\kappa$ is a constant with respect to the original process, hence a normal density for the variations of the rescaled process, $\tilde{X}$, can be easily defined from (3.7), as

$$\Delta\tilde{X}_t \,\dot{\sim}\, \mathcal{N}_2(\kappa\,\mu(X_{t,\dots,t-\tau^{max}},\theta)\,\Delta t,\ \kappa\,\Sigma(X_{t,\dots,t-\tau^{max}},\theta)\kappa\,\Delta t). \tag{3.9}$$

However, mean and variance depend on the population level, $X_t$, which is not observed; nonetheless, mean and variance can easily be rewritten in terms of $\kappa X_t$. In particular we introduce $\tilde{\mu}(\tilde{X}_{t,\dots,t-\tau^{max}}) = \kappa\,\mu(X_{t,\dots,t-\tau^{max}},\theta)$ and $\tilde{\Sigma}(\tilde{X}_{t,\dots,t-\tau^{max}},\theta) = \kappa\,\Sigma(X_{t,\dots,t-\tau^{max}},\theta)\,\kappa$, where $\tilde{X}_{t,\dots,t-\tau^{max}} = (\tilde{X}_t, \dots, \tilde{X}_{t-\tau^{max}})$ and

$$\tilde{\mu}(\tilde{X}_{t,\dots,t-\tau^{max}}) = \begin{pmatrix} \left(c\,\tilde{w}_t^{(1)} - \tilde{w}_t^{(2)} - \tilde{w}_t^{(4)}\right) \\ \left(\frac{1}{c}\,\tilde{w}_t^{(2)} - \tilde{w}_t^{(1)} + \tilde{w}_t^{(3)} - \tilde{w}_t^{(5)}\right) \end{pmatrix} \tag{3.10}$$

and

$$\tilde{\Sigma}(\tilde{X}_{t,\dots,t-\tau^{max}},\theta) = \begin{pmatrix} \kappa_N\left(c\,\tilde{w}_t^{(1)} + \tilde{w}_t^{(2)} + \tilde{w}_t^{(4)}\right) & -\left(\kappa_N\,\tilde{w}_t^{(1)} + \kappa_C\,\tilde{w}_t^{(2)}\right) \\ -\left(\kappa_N\,\tilde{w}_t^{(1)} + \kappa_C\,\tilde{w}_t^{(2)}\right) & \kappa_C\left(\tilde{w}_t^{(1)} + \frac{1}{c}\,\tilde{w}_t^{(2)} + \tilde{w}_t^{(3)} + \tilde{w}_t^{(5)}\right) \end{pmatrix}, \tag{3.11}$$

with $c = \dfrac{\kappa_N}{\kappa_C} = \dfrac{\Omega_N}{\Omega_C}$, the ratio between cytoplasmic and nuclear volumes, and

$$\tilde{w}_t^{(1)} = \qquad \tilde{w}_t^{(1)}(\tilde{X}_t,\theta) \qquad = k_d\,\tilde{X}_t^C \tag{3.12}$$

$$\tilde{w}_t^{(2)} = \tilde{w}_t^{(2)}(\tilde{X}_{t,\dots,t-\tau^{max}}, \theta) = E_\tau(\tilde{X}_{t-\tau}^N)\frac{k_a\,\tilde{X}_t^N}{\kappa_N K_a + \tilde{X}_t^N} \tag{3.13}$$

$$\tilde{w}_t^{(3)} = \tilde{w}_t^{(3)}(\tilde{X}_t, \theta) = \delta\,\kappa_C \tag{3.14}$$

$$\tilde{w}_t^{(4)} = \tilde{w}_t^{(4)}(\tilde{X}_t, \theta) = \gamma\,\tilde{X}_t^N \tag{3.15}$$

$$\tilde{w}_t^{(5)} = \tilde{w}_t^{(5)}(\tilde{X}_t, \theta) = \gamma\,\tilde{X}_t^C. \tag{3.16}$$

The need for the volumes ratio $c$ is due to the fact that the concentrations are expressed with respect to different volumes and one needs to translate the distinct measurement units. It becomes intuitively clear that, if we consider intracellular movements of Nrf2, a decrease of a unit of cytoplasmic concentration in the cytoplasm translates into the increase of a unit of cytoplasmic concentration in the nucleus, which corresponds to an increase of $\frac{\Omega^C}{\Omega^N}$ in units of nuclear concentration.

The three constants in $\kappa$, that is $\tilde{\kappa}$, $\Omega_N$ and $\Omega_C$, only appear in (3.9) in two linearly dependent terms and hence are structurally not identifiable together; in particular, we can only identify two of the three. Although, we could focus on any two combinations of the three, for convenience, we decide to work with parameters $\kappa_N$ and $c$, while $\kappa_C$ is obtained as $\kappa_C = \frac{\kappa_N}{c}$. This parametrization is particularly appealing since, as shown in Section 4.1, it is possible to obtain prior information about $c$ and, hence to formulate an informative prior for this parameter. We redefine our parameter vector including these two constants as $\theta = (k_d, k_a, K_a, \mu_\tau, \sigma_\tau, \gamma, \delta, c, \kappa_N)$.

From (3.9), it is possible to reformulate the DA approximate likelihood in terms of $\tilde{X}$ as

$$L_{DA}(\theta; \tilde{x}) = \prod_{t=1}^{T-1} \Phi(\tilde{x}_{t+1} - \tilde{x}_t|\,\Delta\,\tilde{\mu}(\tilde{x}_{t,\dots,t-\tau^{max}}, \theta), \Delta\,\tilde{\Sigma}(\tilde{x}_{t,\dots,t-\tau^{max}}, \theta)), \tag{3.17}$$

where $\tilde{x} = (\tilde{x}_1, ..., \tilde{x}_T)$ represents the vector of realizations from the random vector $\tilde{X} = (\tilde{X}_1, ..., \tilde{X}_T)$ and, as for $x_{t,\dots,t-\tau^{max}}$, we define $\tilde{x}_{t,\dots,t-\tau^{max}} = (\tilde{x}_t, ..., \tilde{x}_{t-\tau^{max}})$.

## 3.3 Measurement equation and HMM

Further to the proportionality constant, the measurement process is also characterized by a source of white noise which, at time $t$, is assumed to be $\epsilon_t \sim \mathcal{N}_2(0, \Sigma_\epsilon)$ independent and identically distributed (iid) $\forall t$, where $\epsilon_t = (\epsilon_t^N, \epsilon_t^C)^T$ represents the vector of nuclear and cytoplasmic measurement errors at time $t$. In the most general

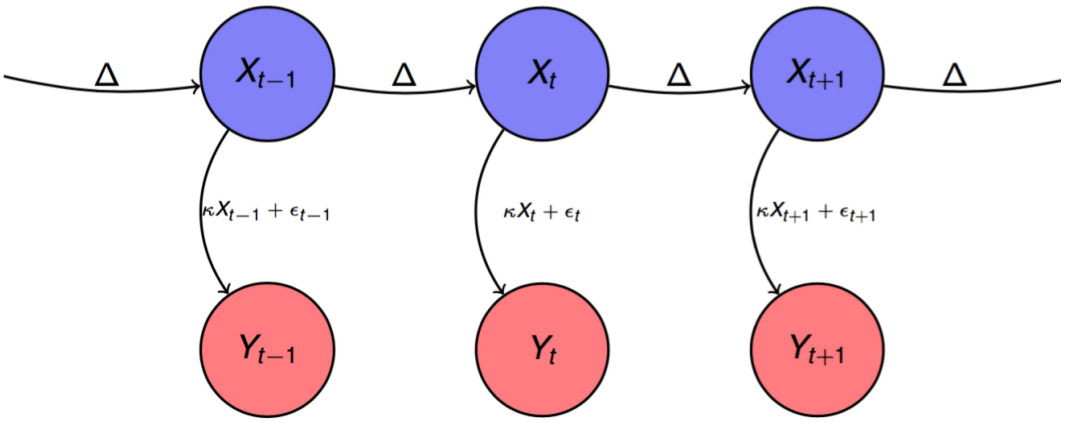


**Figure 3.1:** Latent original process, $X$, and observed one, $Y$.

case, the white noise covariance is defined as:

$$\Sigma_\epsilon = \begin{pmatrix} \sigma_N^2 & \sigma_{NC} \\ \sigma_{NC} & \sigma_C^2 \end{pmatrix}. \tag{3.18}$$

However, in practice, we assume $\sigma_{NC} = 0$. In fact, in an exploratory study on the measurement error, in Section 4.3, as well as in residual analyses, in Section 4.7, we do not detect any significant correlation between nuclear and cytoplasmic error components.

Therefore, the measurement equation relating the original process, $X_t = (X_t^N, X_t^C)^T$, to the observed values, $Y_t = (Y_t^N, Y_t^C)^T$, is

$$Y_t = \kappa X_t + \epsilon_t = \tilde{X}_t + \epsilon_t, \text{ for } t = 1, ..., T, \tag{3.19}$$

where $Y_t^N$ and $Y_t^C$ represent the nuclear and cytoplasmic observations at time $t$. Figure 3.1 graphically illustrates the measurement equation.

Since $Y_t - \tilde{X} = \epsilon_t \sim \mathcal{N}_2(0, \Sigma_\epsilon)$, it is straightforward to formulate a likelihood for the observations as

$$L_\epsilon(\tilde{x}, \Sigma_\epsilon; y) = \prod_{t=1}^{T} \Phi(y_t - \tilde{x}_t | (0,0)^T, \Sigma_\epsilon), \tag{3.20}$$

where $y = (y_1, ..., y_T)$ represents the vector of realizations from the random vector $Y = (Y_1, ..., Y_T)$.

The presence of the measurement error complicates inference; in fact the

original concentration levels are treated as a bivariate continuous latent process: in particular, we deal with a hidden Markov model (HMM) (Baum and Petrie, 1966). The methodology we propose lies in the Bayesian framework and, via a data augmentation procedure (Lesaffre and Lawson, 2012), allows us to reconstruct the unobserved process at discrete time points. In particular, the latent process is treated like a parameter vector and is sampled at the same time points as $Y$ is observed, $t = 1, ..., T$. Clearly, this greatly increases the dimensionality of the posterior space, which is now composed of both parameters and a bivariate latent process with $2\,T$ unobserved points. For simplicity and to decrease the dependence between parameters, we sample the latent process $\tilde{X}$ instead of $X$; however one can easily obtain the latter by multiplying the former by $\kappa^{-1}$, where $\kappa$ has to be estimated.

Inference is implemented via a Metropolis-within-Gibbs algorithm (Metropolis and Ulam, 1949; Metropolis *et al.*, 1953; Hastings, 1970) by alternately sampling from the conditional distributions of the parameters $\theta$ and the latent states $\tilde{X}$. The entire sampling scheme will be described in Section 3.6, once hierarchical structure and priors have been introduced.

## 3.4 Hierarchical modelling

Under the basal and the stimulated conditions, 35 and 36 single cell bivariate processes are observed, respectively. Our study aims at both inferring the kinetic parameters, and obtaining an insight into how they vary between cells. Therefore our goal is to build a model able to distinguish between the intrinsic noise, which refers to the biological birth and death process, and the cell-to-cell variability (Wilkinson, 2012). A natural choice to deal with multiple single cell data is to use a Bayesian hierarchical structure, as in Finkenstädt *et al.* (2013).

Our interest lies in estimating the kinetic parameters in order to get an insight into the cellular mechanism of Nrf2 regulation and improve the accuracy of further simulation studies on this model. We are also interested in understanding the effect that SFN stimulant has on the system. A hierarchical model allows us to explicitly account for the variations between cells, even under the same experimental condition, and to summarize the information from all cells in hyperparameters. Therefore, a hierarchical model is particularly useful for comparing inference, via the hyper-parameters, under the two conditions, allowing us to understand the effect stimulation has on the system, without confounding factors such as the experimental variability.

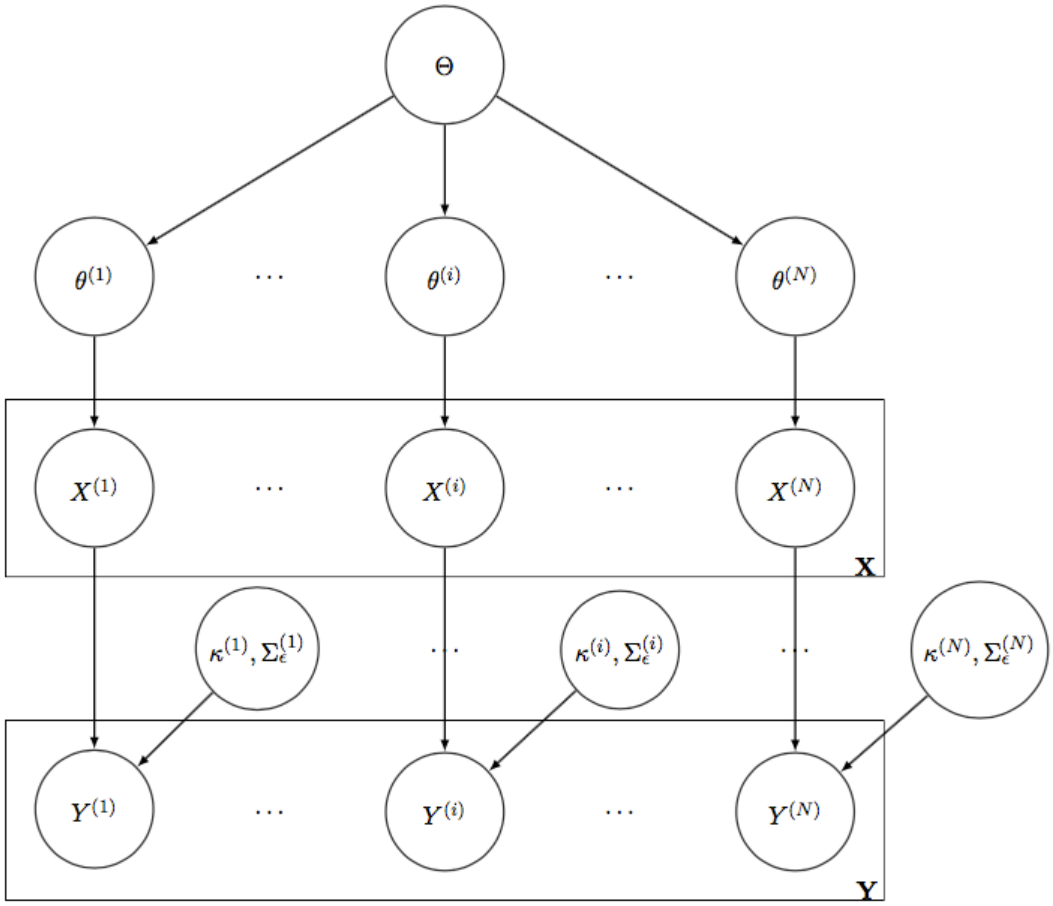


**Figure 3.2:** Graphical model for the hierarchical Nrf2 system. At the top we see the hyperparameters $\Theta$: these generate the hierarchical parameters $\theta^{(1)}, \ldots, \theta^{(N)}$; from the latter ones, the latent processes $X^{(1)}, \ldots, X^{(N)}$ are sampled. Finally, from $X^{(1)}, \ldots, X^{(N)}$ and the measurement equation parameters, we obtain the observed processes $Y^{(1)}, \ldots, Y^{(N)}$.

### Hierarchical likelihood

Let $\mathbf{Y} = (Y^{(1)}, ..., Y^{(N)})$ denote the full data, where $Y^{(i)} = (Y_1^{(i)}, ..., Y_{T_i}^{(i)})^T$ represents the data available for the $i$-th cell, $i = 1, ..., N$, and $T_i$ indicates the total number of observations available for cell $i$, which typically differ from cell to cell, and time is expressed per $\Delta = 2$ minutes, i.e. the time distance between any two consecutive observations $Y_t^{(i)}$ and $Y_{t+1}^{(i)}$ is of 2 minutes.

The measurement equation, relating observations $\mathbf{Y}$ to the original levels $\mathbf{X} = (X^{(1)}, ..., X^{(N)})$, with $X^{(i)} = (X_1^{(i)}, ..., X_{T_i}^{(i)})^T$, now becomes

$$Y_t^{(i)} = \kappa^{(i)} X_t^{(i)} + \epsilon_t^{(i)} = \tilde{X}_t^{(i)} + \epsilon_t^{(i)}, \text{ for } t = 1, ..., T_i \text{ and } i = 1, ..., N, \tag{3.21}$$

with $\epsilon_t^{(i)} \sim \mathcal{N}_2(0, \Sigma_\epsilon^{(i)})$ and where $Y_t^{(i)} = (Y_t^{N(i)}, Y_t^{C(i)})^T$ and $\tilde{X}_t^{(i)} = (\tilde{X}_t^{N(i)}, \tilde{X}_t^{C(i)})^T$, with superscripts $N$ and $C$ indicating the nuclear and cytoplasmic components, respectively. Figure 3.2 visually illustrates the graphical model describing this hierarchical system. We also define the latent process, including the proportionality constant and the volumes, as $\tilde{\mathbf{X}} = (\tilde{X}^{(1)}, ..., \tilde{X}^{(N)})$, where $\tilde{X}^{(i)} = (\tilde{X}_1^{(i)}, ..., \tilde{X}_{T_i}^{(i)})^T = (\kappa^{(i)} X_1^{(i)}, ..., \kappa^{(i)} X_{T_i}^{(i)})^T$. As will be explained in detail in the next Section, for the $i$-th cell, we reconstruct the unobserved process, $\tilde{X}^{(i)}$, and infer the parameter vector $\theta^{(i)}$, by alternately sampling from their conditional distributions.

To keep notation simple, we redefine $\theta^{(i)}$ as the set of all parameters for the $i$-th cell,

$$\theta^{(i)} = (k_d^{(i)}, k_a^{(i)}, K_a^{(i)}, \mu_\tau^{(i)}, \sigma_\tau^{(i)}, \gamma^{(i)}, \delta^{(i)}, c^{(i)}, \kappa_N^{(i)}, \sigma_N^{(i)}, \sigma_C^{(i)})^T, \tag{3.22}$$

including the measurement error ones as well.

Since replicates refer to different cells, they are conditionally independent and, hence, we express the full likelihood, as the product of the individual likelihoods of each component,

$$L_\epsilon(\tilde{\mathbf{X}}, \theta; \mathbf{Y}) = \prod_{i=1}^{N} L_\epsilon(\tilde{x}^{(i)}, \theta^{(i)}; y^{(i)}). \tag{3.23}$$

Analogously we redefine the DA unnormalized density in (3.17) for the full data as

$$L_{DA}(\theta; \tilde{\mathbf{X}}) = \prod_{i=1}^{N} L_{DA}(\theta^{(i)}; \tilde{x}^{(i)}); \tag{3.24}$$

with $\theta = (\theta^{(1)}, ..., \theta^{(N)})$ and where $\theta^{(i)}$, $L_\epsilon(\tilde{x}^{(i)}, \theta^{(i)}; y^{(i)})$ and $L_{DA}(\theta^{(i)}; \tilde{x}^{(i)})$ represent the $i$-th cell parameter vector, likelihood and DA unnormalized density function, defined in (3.20) and (3.17), with $y^{(i)} = (y_1^{(i)}, ..., y_{T_i}^{(i)})^T$ and $\tilde{x}^{(i)} = \tilde{x}^{(i)} = (\tilde{x}_1^{(i)}, ..., \tilde{x}_{T_i}^{(i)})^T$ indicating realizations of the random elements $Y^{(i)}$ and $\tilde{X}^{(i)}$, respectively. In the formulae of this Chapter, the superscript between parentheses refers to the cell index, $i = 1, ..., N$, while the subscript indicates the parameter, $j = 1, ..., p$.

Following the hierarchical paradigm, $\theta^{(i)}$ differs from cell to cell and is assumed to have been generated from a common probability distribution,

$$\theta^{(i)} \sim p(.|\Theta) \, \forall i = 1, ..., N,$$

with $\Theta = (\Theta_1, ..., \Theta_p)$, where each $\Theta_j = (\mu_j, \tau_j)^T$ is the hyperparameter vector quantifying the mean and precision of $\theta_j^{(i)}$ over the cells population. We assume

the prior follows a log-normal distribution, $p(\theta_j^{(i)}|\Theta) \sim log\mathcal{N}(\mu_j, \frac{1}{\tau_j})$, which is a convenient choice as all parameters lie in $\mathbb{R}^+$. Assuming independence in the prior distributions between parameters and cells, the prior density of $\theta$, conditional on hyperparameter $\Theta$, is defined as the double product over the $N$ cells and the $p$ elements in $\theta^{(i)} = (\theta_1^{(i)}, ..., \theta_p^{(i)})^T$,

$$p(\theta|\Theta) = \prod_{i=1}^{N} p(\theta^{(i)}|\Theta) = \prod_{i=1}^{N}\prod_{j=1}^{p} p(\theta_j^{(i)}|\Theta) = \prod_{i=1}^{N}\prod_{j=1}^{p} p(\theta_j^{(i)}|\Theta_j). \tag{3.25}$$

In the hierarchical context, we wish to infer the posterior distribution

$$\begin{aligned} p(\Theta|\mathbf{Y}) &= \int\int p(\theta, \Theta, \tilde{\mathbf{X}}|\mathbf{Y})\, d\theta\, d\tilde{\mathbf{X}} &(3.26)\\ &\propto \int\int L_\epsilon(\tilde{\mathbf{X}}, \theta; \mathbf{Y}) L_{DA}(\theta; \tilde{\mathbf{X}})\, p(\theta|\Theta)\, p(\Theta)\, d\theta\, d\tilde{\mathbf{X}}, &(3.27)\end{aligned}$$

where $p(\Theta)$, usually referred to as hyperprior, denotes the prior of the hyperparameter $\Theta$.

## 3.5 Prior and proposal specifications

In order to carry out our inference, we employ Markov chain Monte Carlo (MCMC) methods; in particular we use a Metropolis-within-Gibbs algorithm (Metropolis and Ulam, 1949; Metropolis *et al.*, 1953; Hastings, 1970).

We use a conjugate normal-gamma hyperprior for each $\Theta_j$, $j = 1, ...p$, $\mu_j|\tau_j \sim \mathcal{N}(\mu_{j_0}, \frac{\lambda_{j_0}}{\tau_j})$ and $\tau_j \sim \mathcal{G}(\alpha_{j_0}, \beta_{j_0})$, where the former denotes the normal distribution with mean $\mu_{j_0}$ and variance $\frac{\lambda_{j_0}}{\tau_j}$ and the latter indicates the gamma random variable (r.v.) with shape and rate parameters $\alpha_{j_0}$ and $\beta_{j_0}$, respectively, i.e. with mean $\frac{\alpha_{j_0}}{\beta_{j_0}}$ and variance $\frac{\alpha_{j_0}}{\beta_{j_0}^2}$. The choice of a conjugate hyperdistribution means that the hypermean and hyperprecision, conditional on the hierarchical parameters, still are normal and gamma distributed, respectively, and hence can be sampled via a Gibbs step, thus decreasing the computational burden.

With the exception of $\delta^{(i)}$, $c^{(i)}$, $\sigma_N^{(i)}$ and $\sigma_C^{(i)}$, on which prior information is available, the hyperpriors were set to be non-informative for all the remaining parameters, with $\mu_{j_0} = 0$, $\lambda_{j_0} = 10^4$, $\alpha_{j_0} = 0.001$ and $\beta_{j_0} = 0.001$. These standard choices correspond to a vague normal prior for the hypermean $\mu_j|\tau_j \sim \mathcal{N}(0, \frac{10^4}{\tau_j})$ and to a vague gamma prior for the hyperprecision $\tau_j$ with mean 1 and variance

$10^3$. The latter is a usual vague prior for the precision parameter, or analogously its inverse for the variance (Gelman, 2006). Such a non-informative prior choice for the precision is used for all parameters: even when prior information is available, we only formulate an informative prior for the hypermean parameter.

### Informative priors and simplifications

We formulate informative hyperpriors for the hypermean parameters of $\delta^{(i)}$, $c^{(i)}$, $\sigma_N^{(i)}$ and $\sigma_C^{(i)}$.

For $\sigma_N^{(i)}$ and $\sigma_C^{(i)}$, the prior information is obtained from an exploratory study on repeated measurements on three cells, which will be described in Section 4.3. In particular, $\mu_{j_0} = 4.41$ and $\lambda_{j_0} = 0.1$ for the nuclear standard deviation, and $\mu_{j_0} = 4.52$, $\lambda_{j_0} = 0.1$, for the cytoplasmic one.

Prior information on the degradation rate was taken from Boisvert et al. (2012). The authors estimate the 50% turnover of Nrf2 proteins to be 5.09 hours, where the 50% turnover is the time until 50% of the original population, which was present at time 0, has changed. Under steady-state conditions, the 50% turnover represents an accurate approximation of the protein half-life, that we call $t_{1/2}$, which is the time until half of the initial population is degraded, assuming no synthesis (Claydon and Beynon, 2012). Under exponential decay, the half-life can easily be converted into the degradation rate. Assuming no synthesis and a constant degradation rate $\delta$ per element of the population, which at time $t$ we call $W_t$, we obtain the following differential equation (DE) for the evolution of $W_t$: $\dfrac{dW_t}{dt} = -\delta\, W_t$. This DE has solution $W_t = W_0\, e^{-\delta\, t}$; from this solution it is possible to express the degradation rate with respect to the half-life, $t_{1/2}$, by substituting $W_{t_{1/2}} = \dfrac{1}{2} W_0$ in the DE solution, we obtain $\delta = \dfrac{ln(2)}{t_{1/2}}$. Hence, by replacing $t_{1/2}$ with its estimate of 305.4 minutes, corresponding to 5.09 hours, we obtain a per minute degradation rate of 0.002269. Therefore we set the degradation hypermean parameters to $\mu_{j_0} = log(0.002269) = -6.088$ and $\lambda_{j_0} = 1$; the choice of $\lambda_{j_0}$, less informative than for measurement error standard deviation, reflects a higher degree of uncertainty in this piece of prior information.

An exploratory study on the ratio of cytoplasmic and nuclear areas, which will be illustrated in Section 4.1, allows us to formulate two informative hyperpriors, one for each condition, for the hypermean of $c^{(i)}$; in particular, we set $\mu_{j_0} = 2.64$ and $\lambda_{j_0} = 0.1$, for the basal condition, and $\mu_{j_0} = 2.47$ and $\lambda_{j_0} = 0.1$, for the stimulated one.

Finally, in simulations, we notice that the standard deviation of the delay

distribution, $\sigma_\tau^{(i)}$, although structurally identifiable, suffers from a lack of practical identifiability, due to the complexity of the model and the limited data available. In order to circumvent this problem, we decide to keep the distributed structure of the delay, which is a more realistic assumption, yet with a fixed standard deviation throughout. Therefore, to decrease the model complexity, $\sigma_\tau$ is chosen not to be hierarchical: $\sigma_\tau^{(i)} = \sigma_\tau, \forall i$. This implies that the delay distribution has a different mean in each cell, although the same variance. After analysing the behaviour of the distribution of $\tau$ for several values of $\sigma_\tau$, we set the standard deviation of the delay $\sigma_\tau = 3$ for all cells.

Therefore we redefine the hierarchical parameter vector we want to infer as $\theta^{(i)} = (k_d^{(i)}, k_a^{(i)}, K_a^{(i)}, \mu_\tau^{(i)}, \gamma^{(i)}, \delta^{(i)}, c^{(i)}, \kappa_N^{(i)}, \sigma_N^{(i)}, \sigma_C^{(i)})^T$

### Adaptive random walk proposal

The sampling of the hierarchical parameters in $\theta^{(i)}$ follows a Metropolis-within-Gibbs scheme, where movements for each $\theta^{(i)}$ are proposed and accepted in five blocks, that we define as, $\theta_{(b_1)}^{(i)} = (k_d^{(i)}, \mu_\tau^{(i)})$, $\theta_{(b_2)}^{(i)} = (k_a^{(i)}, K_a^{(i)})$, $\theta_{(b_3)}^{(i)} = (\delta^{(i)}, \gamma^{(i)})$, $\theta_{(b_4)}^{(i)} = (c^{(i)}, \kappa_N^{(i)})$ and $\theta_{(b_5)}^{(i)} = (\sigma_N^{(i)}, \sigma_C^{(i)})$. The blocks are chosen, after an initial analysis where each hierarchical parameter is proposed independently from a simple random walk (RW), by merging, in the same block, the parameters with the most correlated posterior chains. We also define $b_1 = \{1, 4\}$, $b_2 = \{2, 3\}$, $b_3 = \{5, 6\}$, $b_4 = \{7, 8\}$ and $b_5 = \{9, 10\}$ as the vectors indicating the elements of $\theta^{(i)}$ belonging to each one of the five blocks.

For each $i$, proposals in each block are sampled, in the log space, according to the adaptive random walk (ARW) scheme (Haario *et al.*, 2001), from a normal distribution centred around the previous iteration values with variance proportional to the covariance matrix estimated from the parameter chains of the respective block. The adaptation is analogous to the one implemented by Haario *et al.* (2001), where constants $\epsilon$ and $s_d$ are chosen in order to optimize each block's acceptance rate. The MCMC is first run for 2,000 iterations without adaptation, as a standard random walk (RW), and only then the covariance matrices are computed from the chains, excluding the first 1,000 values, and they are used to tune the proposal variance. Being the correlation computed on all values of the chain from a fixed starting point onwards, the diminishing adaptation requirement (Roberts and Rosenthal, 2009) is respected. In other words, the proposal distribution stabilises as the chains increase; i.e. the influence, on the proposal distribution, of the $r$-th iteration of the MCMC, goes to 0 as $r \to \infty$.

## 3.6 Metropolis-within-Gibbs algorithm

This Section describes in detail the sampling scheme for all our parameters, $\Theta$, $\theta$ and $\tilde{\mathbf{X}}$. Since all parameters in $\theta^{(i)}$ are positive, we sample them in the log space, where the prior for $log(\theta_j^{(i)})$ is $p(log(\theta_j^{(i)})|\Theta) \sim \mathcal{N}(\mu_j, \frac{1}{\tau_j})$, for each $i = 1, ..., N$ and $j = 1, ..., p$. Inference is implemented via a Metropolis-within-Gibbs algorithm (Metropolis and Ulam, 1949; Metropolis *et al.*, 1953; Hastings, 1970) through the following steps.

**Initialisation**: We initialize $\tilde{\mathbf{X}}$ equal to $\mathbf{Y}$; we also set the parameters in $\Theta$ and $\theta$ to arbitrary real positive values. In the real data analysis $\theta$ is chosen after an initial exploratory data analysis.

**Update**: At each iteration of the MCMC, $r = 1, ..., R$, where $R$ represents the total number of iterations the algorithm is run for, we conditionally update the parameters by iteratively repeating the following steps.

Update $\Theta|\theta$: For every $j = 1, ..., p$ independently, each hyper-parameter $\Theta_j = (\mu_j, \tau_j)^T$ is sampled, conditional on the corresponding hierarchical parameters $\theta_j = (\theta_j^{(1)}, ..., \theta_j^{(N)})$, from a Gibbs step, thanks to the conjugacy of the hyper-priors. In particular the hypermean *a posteriori* still is normal,

$$\mu_j|\tau_j \sim \mathcal{N}\left(\frac{\tau_j/\lambda_{j_0}\,\mu_{j_0} + \tau_j \sum_{i=1}^{N} log(\theta_j^{(i)})}{\tau_j/\lambda_{j_0} + N\,\tau_j}, (\tau_j/\lambda_{j_0} + N\,\tau_j)^{-1}\right),$$

with $\mathcal{N}(\mu, \sigma^2)$ being the univariate normal distribution with mean $\mu$ and variance $\sigma^2$, and the hyperprecision is again *a posteriori* gamma distributed (Gelman, 2006),

$$\tau_j|\mu_j \sim Gamma\left(\alpha_{j_0} + N/2, \beta_{j_0} + 1/2 \sum_{i=1}^{N} \left(log(\theta_j^{(i)}) - \mu_j\right)^2\right),$$

where $Gamma(\alpha, \beta)$ denotes the gamma r.v. with shape and rate parameters $\alpha$ and $\beta$, respectively, i.e. with mean $\frac{\alpha}{\beta}$ and variance $\frac{\alpha}{\beta^2}$.

Update $\theta|\mathbf{Y}, \tilde{\mathbf{X}}, \Theta$: Parameters in $log(\theta^{(i)})$ are sampled, independently in every cell $i = 1, ..., N$, from an adaptive Metropolis step.

Every $log(\theta^{(i)})$ is sampled in 5 blocks, as described above, and candidates are proposed from a normal distribution with an adaptive covariance matrix which, for the $i$-th cell and $j$-th block, we call $\Sigma_{ARW_j}^{(i)}$.

For each block $j = 1, ..., 5$, $log(\tilde{\theta}^{(i)}_{(b_j)})$ is sampled from the bivariate normal $\mathcal{N}_2(log(\theta^{(i)}_{(b_j)}), \Sigma^{(i)}_{ARW_j})$, where $\theta^{(i)}_{(b_j)}$ and $\tilde{\theta}^{(i)}_{(b_j)}$ represent the current iteration and the proposed parameter values for block $j$, respectively.

For blocks $j = 1, ..., 4$, each $\tilde{\theta}^{(i)}_{b_j}$ is accepted with probability $min(1, A^{(i)}_j)$, with

$$A^{(i)}_j = \frac{L_{DA}((\theta^{(i)}_{-(b_j)}, \tilde{\theta}^{(i)}_{(b_j)}); \tilde{x}^{(i)}) \prod_{j \in b_j} \Phi(log(\tilde{\theta}^{(i)}_j); \mu_j, \frac{1}{\tau_j})}{L_{DA}(\theta^{(i)}; \tilde{x}^{(i)}) \prod_{j \in b_j} \Phi(log(\theta^{(i)}_j); \mu_j, \frac{1}{\tau_j})},$$

where $(\theta^{(i)}_{-(b_j)}, \tilde{\theta}^{(i)}_{(b_j)})$ represents the parameter vector with elements for block $j$ from $\tilde{\theta}^{(i)}$ and for all other blocks from $\theta^{(i)}$.

Block 5 follows a different acceptance probability instead, as it refers to the measurement error standard deviations; in particular $\tilde{\theta}^{(i)}_{(b_5)}$ is accepted with probability $min(1, A^{(i)}_5)$, with

$$A^{(i)}_5 = \frac{L_\epsilon(\tilde{x}^{(i)}, \tilde{\theta}^{(i)}_{(b_5)}; y^{(i)}) \prod_{j \in b_5} \Phi(log(\tilde{\theta}^{(i)}_j); \mu_j, \frac{1}{\tau_j})}{L_\epsilon(\tilde{x}^{(i)}, \theta^{(i)}_{(b_5)}; y^{(i)}) \prod_{j \in b_5} \Phi(log(\theta^{(i)}_j); \mu_j, \frac{1}{\tau_j})},$$

Update $\tilde{\mathbf{X}}|\mathbf{Y}, \theta$: For each cell, $i = 1, ..., N$, we sample, independently, the latent process $\tilde{X}^{(i)}$ conditional on parameters $\theta^{(i)}$ and on observations $Y^{(i)}$. Each time point $\tilde{X}^{(i)}_t$, with $t = 1, ..., T_i$, is updated conditional on the rest of the process: we start by sampling the latent data at time 1 and proceed, one step at a time, until time $T_i$.

* For $t = 2, ..., T_i - 1$, $\tilde{X}^{(i)}_t$ is sampled, conditional on the remaining latent states, which we call $\tilde{X}^{(i)}_{-(t)} = \tilde{x}^{(i)}_{-(t)}$, targeting density

$$\begin{aligned} f(\tilde{x}^{(i)}_t | \tilde{x}^{(i)}_{-(t)}, y^{(i)}_t, \theta^{(i)}) \quad \propto \quad & \Phi(\tilde{x}^{(i)}_t - \tilde{x}^{(i)}_{t-1} | \mu_{t-1}, \Sigma_{t-1}) \times \\ & \Phi(\tilde{x}^{(i)}_{t+1} - \tilde{x}^{(i)}_t | \mu_t, \Sigma_t) \times \\ & \Phi(y^{(i)}_t - \tilde{x}^{(i)}_t | (0,0)^T, \Sigma_\epsilon), \end{aligned} \tag{3.28}$$

where $\mu_t = \Delta\tilde{\mu}(\tilde{x}_{t,...,t-\tau^{max}}, \theta)$ and $\Sigma_t = \Delta\tilde{\Sigma}(\tilde{x}_{t,...,t-\tau^{max}}, \theta))$ are the mean and variance of a variation of the process from time $t$ to time $t+1$ from the DA in (3.17) with $\Delta$ representing the time interval between consecutive observations, which is 2 minutes in our case. The first two terms on the right hand side (RHS) of the equation represent the densities of the changes from $\tilde{x}^{(i)}_{t-1}$ to $\tilde{x}^{(i)}_t$ and from $\tilde{x}^{(i)}_t$

to $\tilde{x}_{t+1}^{(i)}$, that is the two variations in the latent process, in intervals of length $\Delta$, before and after $\tilde{x}_t^{(i)}$. The last term of the RHS derives from the fact that $y_t^{(i)} - \tilde{x}_t^{(i)}$ represents the measurement error term, which is distributed like a bivariate normal $\mathcal{N}_2((0,0)^T, \Sigma_\epsilon)$.
It is not possible to introduce a Gibbs step for $\tilde{x}_t^{(i)}$, proposing directly from $f(\tilde{x}_t|\tilde{x}_{-(t)}^{(i)}, y_t^{(i)}, \theta^{(i)})$, as $\tilde{x}_t^{(i)}$ appears non-linearly in the mean and variance of the second density on the RHS of (3.28); therefore the three densities, jointly, are no longer normal. Instead, we employ a Metropolis-Hastings (MH) algorithm targeting $f(\tilde{x}_t|\tilde{x}_{-(t)}^{(i)}, y_t^{(i)}, \theta^{(i)})$. In order to increase the acceptance rate, we look for a proposal distribution as close as possible to the target one. In particular, each $\tilde{x}_t^{(i)}$ is proposed from $q_t(x) \sim \mathcal{N}_2(\mu_{q_t}, \Sigma_{q_t})$, where $\Sigma_{q_t} = \left(\Sigma_\epsilon^{-1} + \Sigma_{t-1}^{-1}\right)^{-1}$ and $\mu_{q_t} = \Sigma_{q_t}\left(\Sigma_\epsilon^{-1}\, y_t^{(i)} + \Sigma_{t-1}^{-1}\,(\tilde{x}_{t-1}^{(i)} + \mu_{t-1})\right)$. The bivariate normal proposal $q(.)$ is obtained as the joint distribution of the first and third components on the RHS of (3.28), which are both normals with parameters that do not depend on $\tilde{x}_t^{(i)}$.
We call $\tilde{w}_t^{(i)}$ the candidate value for the $i$-th cell latent process at time $t$: the acceptance rate for $w_t^{(i)}$ is $min(1, A_t^{(i)})$, with

$$A_t^{(i)} = \frac{f(\tilde{w}_t|\tilde{x}_{-(t)}^{(i)}, y_t^{(i)}, \theta^{(i)})\, q_t(\tilde{x}_t)}{f(\tilde{x}_t|\tilde{x}_{-(t)}^{(i)}, y_t^{(i)}, \theta^{(i)})\, q_t(w)} = \frac{\Phi(\tilde{x}_{t+1}^{(i)} - \tilde{w}_t^{(i)}|\mu_t, \Sigma_t)}{\Phi(\tilde{x}_{t+1}^{(i)} - \tilde{x}_t^{(i)}|\mu_t, \Sigma_t)}.$$

In the acceptance probability, proposal and target densities simplify, leading to an increased acceptance rate, i.e. the fraction of proposed values which are accepted. In the experimental data analysis in Section 4.6, we obtain an average acceptance rate, for each point of the latent process, of about 70%.

∗ For $t = 1$, the target density we sample from is slightly different as the process is unavailable before time 1. In this case $\tilde{x}_1^{(i)}$ is sampled, via a MH step, from

$$f_1(\tilde{x}_1^{(i)}|\tilde{x}_{-(1)}^{(i)}, y_1^{(i)}, \theta^{(i)}) \propto \Phi(\tilde{x}_2^{(i)} - \tilde{x}_1^{(i)}|\mu_1, \Sigma_1) \times \Phi(y_1^{(i)} - \tilde{x}_1^{(i)}|(0,0)^T, \Sigma_\epsilon). \quad (3.29)$$

In this case we propose from the second density on the RHS of (3.29); i.e. we sample $\tilde{w}_1^{(i)}$ from $q_1(x) \sim \mathcal{N}_2(y_1^{(i)}, \Sigma_\epsilon)$.
The acceptance rate for $\tilde{w}_1^{(i)}$ is $min(1, A_1^{(i)})$, where again proposal

and target densities simplify in $A_1^{(i)}$ which becomes

$$A_1^{(i)} = \frac{f_1(\tilde{w}_1^{(i)}|\tilde{x}_{-(1)}^{(i)}, y_1^{(i)}, \theta^{(i)})\, q(\tilde{x}_1^{(i)})}{f_1(\tilde{x}_1^{(i)}|\tilde{x}_{-(1)}^{(i)}, y_1^{(i)}, \theta^{(i)})\, q(\tilde{w}_1^{(i)})} = \frac{\Phi(\tilde{x}_2^{(i)} - \tilde{w}_1^{(i)}|\mu_1, \Sigma_1)}{\Phi(\tilde{x}_2^{(i)} - \tilde{x}_1^{(i)}|\mu_1, \Sigma_1)}$$

* For $t = T_i$ again our target density varies because the process is unavailable after time $T_i$; our target density is

$$\begin{aligned} f_T(\tilde{x}_{T_i}^{(i)}|\tilde{x}_{-(T_i)}^{(i)}, y_t^{(i)}, \theta^{(i)}) \quad &\propto \quad \Phi(\tilde{x}_{T_i}^{(i)} - \tilde{x}_{T_i-1}^{(i)}|\mu_{T_i-1}, \Sigma_{T_i-1}) \times \\ &\Phi(y_{T_i}^{(i)} - \tilde{x}_{T_i}^{(i)}|(0,0)^T, \Sigma_\epsilon). \end{aligned} \tag{3.30}$$

In this case we can implement a Gibbs step and sample $\tilde{x}_{T_i}^{(i)}$ directly from $f_T(.)$ in (3.30), which is a bivariate normal, $f_T(x) \sim \mathcal{N}_2(\mu_{q_{T_i}}, \Sigma_{q_{T_i}})$, with variance and mean defined as $\Sigma_{q_{T_i}} = \left(\Sigma_\epsilon^{-1} + \Sigma_{T_i-1}^{-1}\right)^{-1}$ and $\mu_{q_{T_i}} = \Sigma_{q_{T_i}} \left(\Sigma_\epsilon^{-1}\, y_{T_i}^{(i)} + \Sigma_{T_i-1}^{-1}\, (\tilde{x}_{T_i-1}^{(i)} + \mu_{T_i-1})\right)$, respectively, where all quantities have been defined above when introducing the proposal density $q_t(.)$.

## 3.7 Simulation study

We perform a simulation study to validate the performance of the methodology described in this Chapter. In particular, we run 3 hierarchical simulations independently, each comprising of N = 35 processes of exactly the same length and time resolution, 2 minutes, as the ones available for the cells observed under the basal condition, see Table 3.1, on average 222 bivariate observations are available for each cell.

| Cell | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Observations | 172 | 251 | 265 | 417 | 288 | 210 | 244 | 152 | 154 | 150 |
| Cell | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| Observations | 217 | 218 | 147 | 245 | 111 | 268 | 249 | 345 | 218 | 171 |
| Cell | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
| Observations | 326 | 196 | 255 | 250 | 178 | 196 | 150 | 198 | 198 | 356 |
| Cell | 31 | 32 | 33 | 34 | 35 | | | | | |
| Observations | 284 | 147 | 150 | 184 | 206 | | | | | |

**Table 3.1:** Number of observations available for each cell in each simulation study; the average per cell equals 221.9.

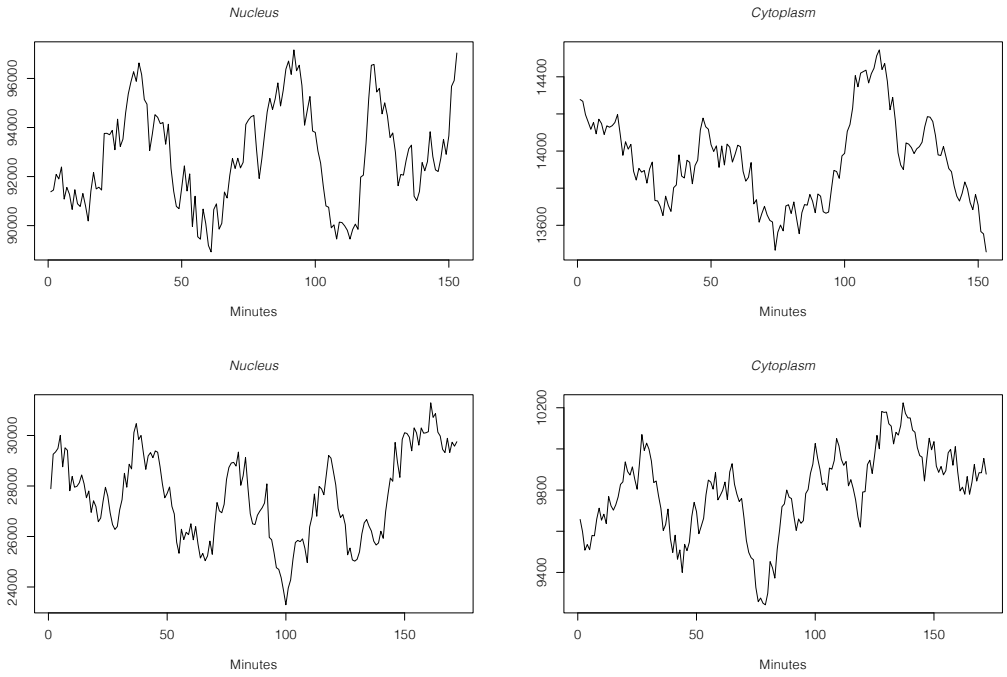


**Figure 3.3:** Example of two simulated processes for nuclear and cytoplasmic observations. Parameters used in the top row simulation are those for cell 8 in Table A.1; the bottom row simulation corresponds to cell 20 from the same table.

Firstly, we choose 3 similar sets of hyperprior values, $\Theta$, and, on each, we sample $N = 35$ hierarchical parameter vectors, $\theta^{(i)} \sim p(.|\Theta)$; i.e. in every simulation we sample each parameter $\theta_j^{(i)}$ from $log\mathcal{N}(\mu_j, \frac{1}{\tau_j})$, $i = 1, ..., N$ and $j = 1, ..., p$. hyperparameters are chosen in order to reproduce similar patterns as the ones exhibited by the available experimental data. Tables A.1-A.3, in the Appendix, report all the simulated hierarchical parameter values in $\theta$, for the three simulation studies.

In each of the three simulations, for every hierarchical parameter vector, $\theta^{(i)}$, we simulate a bivariate process, for nuclear and cytoplasmic observations $Y^{(i)}$. Firstly, we initialize nuclear and cytoplasmic molecular populations, $X_0^{(i)}$, to the average levels observed in the experimental data, under the basal condition, divided by the respective $\kappa^{(i)}$. We then sample the population process at the next time point, $t + 1$, via the Euler-Maruyama approximation; i.e. in each cell, $X_{t+1}^{(i)}$ is sampled from its approximate normal distribution in (3.7), $X_{t+1}^{(i)} \dot{\sim} \mathcal{N}_2(X_t^{(i)} + \mu(X_{t,...,t-\tau^{max}}^{(i)}, \theta^{(i)})\,\Delta t,\ \Sigma(X_{t,...,t-\tau^{max}}^{(i)}, \theta^{(i)})\,\Delta t)$, where $\Delta = 2$ minutes. For the first $\tau^{max} = 60$ minutes the delay component is unavailable and hence is replaced by the starting value $X_0^{(i)}$. Each process is sampled for $1,000$ minutes, which are discarded

in order to remove the effect of the fixed delay used in the first 60 minutes, before simulating the process for further $T_i$ steps, thus obtaining $X^{(i)} = (X_1^{(i)}, ..., X_{T_i}^{(i)})^T$. Following measurement equation (3.21) we multiply each simulated process $X^{(i)}$ by $\kappa^{(i)}$ and add a normal error component $\epsilon^{(i)}$ obtaining, in each cell, simulated data for $Y^{(i)}$.

By repeating this procedure, in each of the three simulations, for every simulated parameter $\theta^{(i)}$, $i = 1, ..., N$, we obtain three simulated datasets $\mathbf{Y}$, one per simulation study.

Figure 3.3 shows, for illustrative purposes, two simulated processes.

### Inferential Results

We perform inference, independently in each of the three simulations, following the algorithm described in Section 3.6.

In each MCMC chain, after discarding at least the first $10^5$ iterations as a *burn-in*, we let the algorithm run for at least $4 \times 10^5$ more iterations, which are used to approximate the posterior chains. Inference is on the overall successful on the hierarchical parameters as well as on the hyper ones.

Figure 3.4 shows the traceplots for logarithm of the overall DA unnormalized density of parameters, $L_{DA}(\theta; \tilde{\mathbf{X}})$ in (3.24), in the three simulation studies.

Table 3.2 displays the empirical coverages of parameters, computed over 105 hierarchical parameters (35 for each simulation), of the 0.99 and 0.95 level highest posterior density (HPD) credible intervals (CIs), computed in  (R Core Team, 2016) via the *HPDinterval* function of package *coda* (Plummer *et al.*, 2016).

Results provide good evidence that our algorithm is a successful inferential tool, especially if considering the overall coverages, averaged over all kinetic parameters, in the last column of Table 3.2. Nonetheless, $k_a$ and $K_a$ seem to have slightly lower coverages than the nominal ones. This is probably due to the high dependence between the two, which appear in likelihood (3.17) via the Michaelis-Menten term,

| Level | $k_d^{(i)}$ | $k_a^{(i)}$ | $K_a^{(i)}$ | $\mu_\tau^{(i)}$ | $\gamma^{(i)}$ | $\delta^{(i)}$ | $c^{(i)}$ | $\kappa_N^{(i)}$ | Tot |
|---|---|---|---|---|---|---|---|---|---|
| 99 | 100.0 | 97.1 | 98.1 | 99.0 | 100.0 | 100.0 | 100.0 | 100.0 | 99.3 |
| 95 | 94.3 | 88.6 | 91.4 | 93.3 | 97.1 | 100.0 | 96.2 | 99.0 | 95.0 |

**Table 3.2:** Empirical coverages, in percentage, of 0.99 and 0.95 level HPD credible intervals, based on 105 hierarchical parameters from 3 simulations, each with 35 simulated processed. The last column indicates the average coverage of parameters ($k_d^{(i)}$, $k_a^{(i)}$, $K_a^{(i)}$, $\mu_\tau^{(i)}$, $\gamma^{(i)}$, $\delta^{(i)}$, $c^{(i)}$, $\kappa_N^{(i)}$).

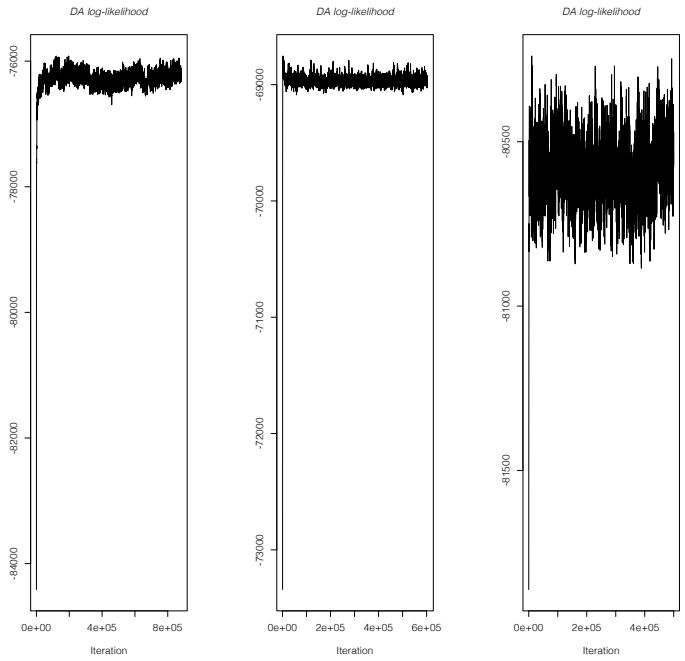


**Figure 3.4:** Chains for the full $L_{DA}$, in (3.24), from left to right, for the first, second and third simulation studies.

$\frac{k_a \tilde{X}_t^N}{\kappa_N K_a + \tilde{X}_t^N}$. Indeed, the CIs for their ratio, $\frac{k_a^{(i)}}{K_a^{(i)}}$, have excellent coverage: 99.9 and 96.6% for the 0.99 and 0.95 level CI respectively. This suggests that, although the two parameters alone may not always be well estimated, their ratio is. Figure 3.5 shows an example of a single cell profile log-likelihood, for parameters $k_a^{(i)}$ and $K_a^{(i)}$, in an individual simulated process, where the rest of parameters are set to their true values. It is interesting to notice that it is hard it is to identify, by eye, a maximum for the likelihood, as the region for the highest values of the log-likelihood seems highly elongated, confirming the strong dependence between the two parameters.

The CIs for the measurement error parameters, in Table 3.3, appear to suffer from some *under-coverage*. As they represent nuissance parameters, we are less worried by this phenomenon, given it does not seem severe and does not have a visible effect on inference on the parameters of interest. We also recall the fact that, in general, Bayesian CIs of level $1 - \alpha$, unlike frequentist confidence intervals, do not have the property of including the parameter of interest with probability $1 - \alpha$ and, instead, represent a posterior belief that the parameter belongs to the interval (Lesaffre and Lawson, 2012).

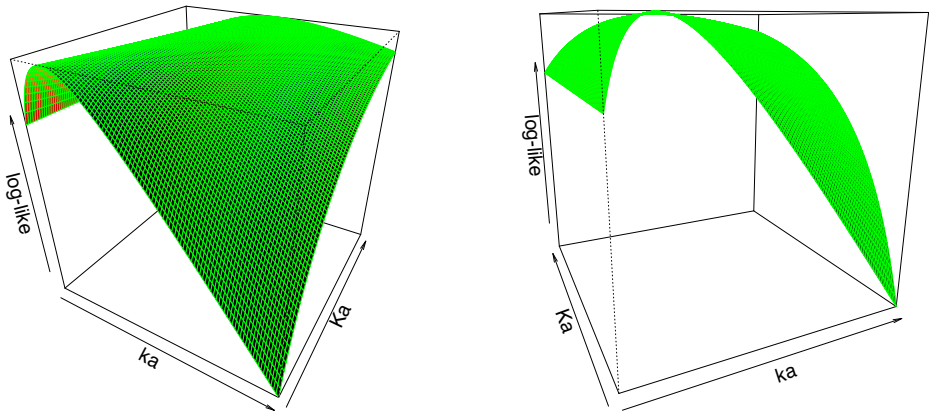


**Figure 3.5:** Example of a 3D plot of the profile log-likelihood for parameters $k_a^{(i)}$ and $K_a^{(i)}$ in a simulated process in a single cell.

| Level | $\sigma_N^{(i)}$ | $\sigma_C^{(i)}$ |
|---|---|---|
| 99 | 95.2 | 97.1 |
| 95 | 89.5 | 89.5 |

**Table 3.3:** Empirical coverages, in percentage, for the measurement error parameters, of 0.99 and 0.95 level HPD credible intervals, based on 105 hierarchical parameters from 4 simulations, each with 35 simulated processed.

We also obtain encouraging results for inference on the hyperparameters: all 30 hypermeans and all 30 hyperprecisions fall inside, both, their 0.99 and 0.95 HPD CIs. Figures B.11-B.13 in the Appendix report the images for the posterior densities of the hypermean parameters in each simulation study. The respective real values used in the simulations, denoted by a vertical line, are always in a central area of the corresponding posterior density.

For illustrative purposes, Figure 3.6 displays an example, for a simulated cell, on how the method reconstructs the unobserved process: the blue dashed lines represent the 0.95 point-wise HPD CIs of the latent process, which is represented by the black solid line.

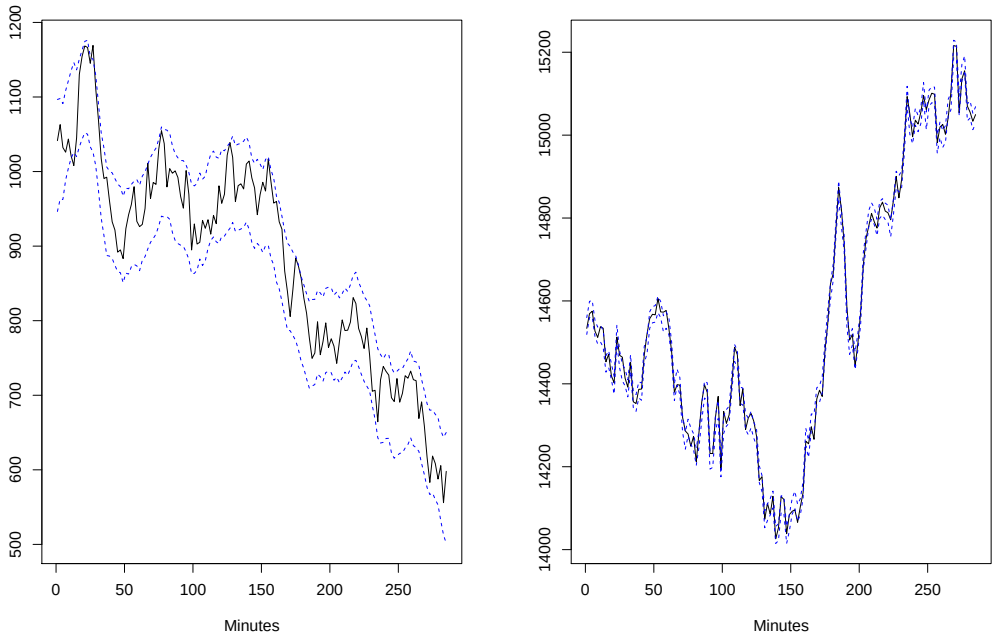


**Figure 3.6:** Example on a cell of the original $\kappa X$ process, black solid line, and respective 0.95 HPD CI, blue dashed lines, for nuclear (left) and cytoplasmic (right) levels.

## 3.8 Summary

In this Chapter we have introduced a methodology to infer parameters of a continuous time stochastic system via a Bayesian hierarchical model, from multiple single cell data observed at discrete time points only. Observations are also characterized by a source of random measurement error; this introduces latent states that we deal with via a data augmentation procedure. Inference is carried out, following the Bayesian paradigm, by alternately sampling from the conditional distributions of the hidden states and parameters, with a Metropolis-within-Gibbs algorithm. We validate our method in three independent hierarchical simulation studies which show that the inferential methodology is successful in retrieving the unknown parameters. An alternative approach to the data augmentation could be, for instance, the use of a particle method, such as a particle marginal Metropolis-Hastings (PMMH) algorithm which allows us to estimate the marginal density of the data by integrating with respect to the latent states. We also employed this procedure, but due to its high computational burden, in our case, the posterior chains of parameters needed more time to converge and provide an accurate estimate of the posterior densities,

than the data augmentation procedure.

In the next Chapter, we will first discuss some exploratory data analyses which provide us initial information about specific aspects of our data which, in some cases, will be embedded into an informative prior, as previously described in Section 3.5. Then, the methodology presented here will be applied to the observed Nrf2 protein levels to infer the parameters of its underlying stochastic system.

# Chapter 4

# Experimental data analysis

> Theory is when we know everything but nothing works. Practice is when everything works and nobody knows why. We always end up by combining the two: nothing works and nobody knows why.
>
> *Albert Einstein*

In this Chapter we describe the analyses of the observed data on Nrf2 concentration levels in nucleus and cytoplasm. As explained in Chapter 2, the experimental data is available for multiple cells under two experimental conditions: under the basal condition and under the induction by a stimulant, *sulforaphane* at $2\mu M$. The observed processes are reproduced in Figures B.1-B.10, in the Appendix, and the number of observations for each cell under the two conditions is displayed in Tables 4.1 and 4.2.

| Cell | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Observations | 172 | 251 | 265 | 417 | 288 | 210 | 244 | 152 | 154 | 150 |
| Cell | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| Observations | 217 | 218 | 147 | 245 | 111 | 268 | 249 | 345 | 218 | 171 |
| Cell | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
| Observations | 326 | 196 | 255 | 250 | 178 | 196 | 150 | 198 | 198 | 356 |
| Cell | 31 | 32 | 33 | 34 | 35 | | | | | |
| Observations | 284 | 147 | 150 | 184 | 206 | | | | | |

**Table 4.1:** Number of observations available for each cell in the experimental data under the basal condition; the average number per cell equals 221.9.

| Cell | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Observations | 361 | 199 | 167 | 133 | 237 | 166 | 219 | 168 | 185 | 130 |
| Cell | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| Observations | 348 | 135 | 118 | 189 | 202 | 230 | 230 | 191 | 241 | 150 |
| Cell | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
| Observations | 175 | 179 | 156 | 209 | 209 | 148 | 150 | 184 | 232 | 102 |
| Cell | 31 | 32 | 33 | 34 | 35 | 36 | | | | |
| Observations | 218 | 135 | 200 | 271 | 123 | 269 | | | | |

**Table 4.2:** Number of observations available for each cell in the experimental data under the stimulated condition; the average number per cell equals 193.3.

In the first part of this Chapter, we describe some exploratory data analyses which allow us to get an initial insight into our data and to formulate informative priors for some of our parameters. In the second part, we use the methodology described in the previous Chapter to infer the kinetic parameters of the system from the available data with the aim of comparing the results for the two experimental conditions.

## 4.1 Exploratory analysis on the volumes ratio

In order to gain an insight into the ratio between cytoplasmic and nuclear volumes, we study the relation between their areas, obtained by the 2-dimensional projections of their structure, as shown in Figure 2.3. We compute, via the *CellTracker* software (Du and Bretschneider, 2013), the areas, expressed in terms of pixels, of nucleus and cell, which we call $A^{(i)}_{Nuc}$ and $A^{(i)}_{Cell}$, respectively, defined by our borders (Figure 2.3). The areas are computed, on the available experimental data, at the first time the cell is observed, which we call time t = 1; from these values we compute the cytoplasmic area, $A^{(i)}_{Cyt} = A^{(i)}_{Cell} - A^{(i)}_{Nuc}$, and the ratio between the cytoplasmic and nuclear areas for each cell, which we define as $A^{(i)} = \dfrac{A^{(i)}_{Cyt}}{A^{(i)}_{Nuc}}$. We distinguish between the ratios estimated for the basal, $A^{(i)}_B$, and stimulated, $A^{(i)}_S$, conditions. In particular, we observe 29 in the former case and 50 in the latter. Numbers do not match the available cells for two main reasons. Firstly, some cells are excluded from the current analysis as they are only visible some time after the start of the experiment; in some cases instead, we exploit further available cells which are not used in the data analysis, because we believe the transient transfection process, described in Section 2.2, might have not properly worked in these cells and, hence, the reporter mRNA

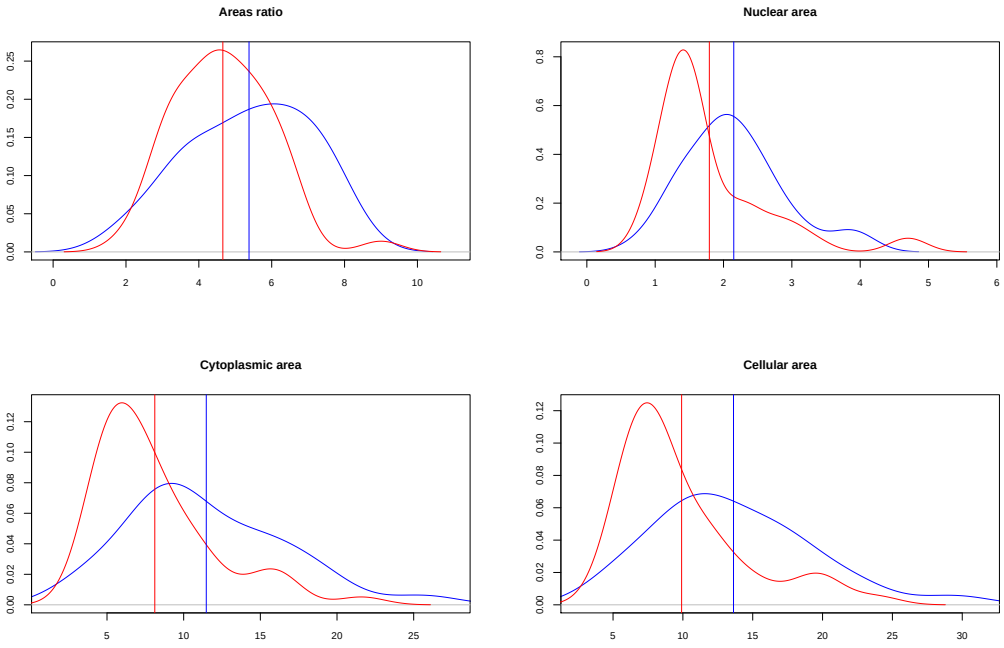


**Figure 4.1:** Densities of the areas, computed at time t = 1, for the cells under the basal condition (in blue) and under the induction of a stimulant (in red). The vertical solid lines denote the mean values.

in these cells may not accurately represent the original mRNA levels. Clearly this does not affect the size of cells, which can be therefore used in the present analysis that is only based on the cellular borders.

Figure 4.1 shows the densities for these ratios as well as for the of the nuclear, cytoplasmic and cellular estimated areas, for cells observed under the basal condition (in blue) and for those under the induction by a stimulant (in red): these images suggest that the stimulant might have an effect on the relation between the nuclear and cytoplasmic areas where the latter seems to be the most influenced out of the two.

In order to investigate whether there is a significant difference in the ratios between basal and stimulated conditions, we perform a permutation test on the mean of this ratio (Fisher, 1935). In particular, assuming $A_B^{(i)}$, $i = 1, ..., 29$, and $A_S^{(i)}$, $i = 1, ..., 50$, are generated from some distribution with mean $\mu_{ratio}^B$, for the basal condition, and $\mu_{ratio}^S$ for the stimulated condition, we test null hypothesis $H_0 : \mu_{ratio}^B = \mu_{ratio}^S$ versus the alternative hypothesis $H_1 : \mu_{ratio}^B \neq \mu_{ratio}^S$. We perform a permutation test on the absolute differences between the observed means

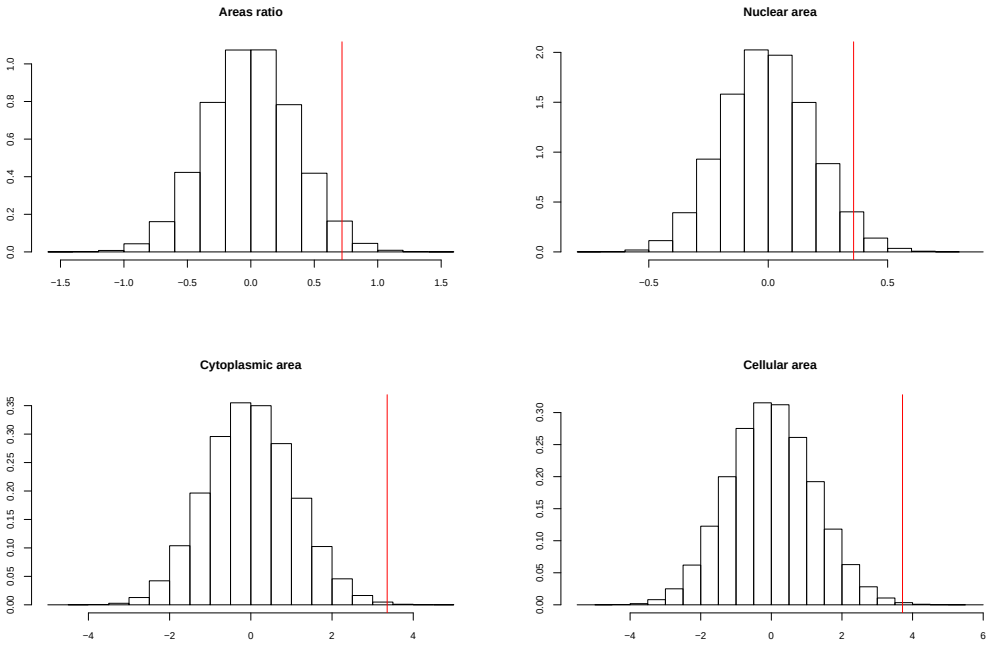


**Figure 4.2:** Null distribution, for the mean difference, of permutation tests on the mean of the ratio of areas and of nuclear, cytoplasmic and cellular areas, at time 1. The vertical solid line represents the observed value of the statistic.

in the two samples, $S = \bar{A}_B - \bar{A}_S$, where $\bar{A}_B = \frac{1}{29}\sum_{i=1}^{29} A_B^{(i)}$ and $\bar{A}_A = \frac{1}{50}\sum_{i=1}^{50} A_S^{(i)}$. First we sample the null distribution of statistic $S$ by randomly permuting, $10^6$ times, the 79 elements of $A = (A_B^{(1)}, ..., A_B^{(29)}, A_S^{(1)}, ..., A_S^{(50)})$ into two groups, of size 29 and 50, and computing the statistic S on each permutation $p = 1, ..., 10^6$, which for permutation $p$ we call $S^p$. In this way we approximate, with $10^6$ samples, the distribution of $S$ under the null hypothesis that the mean of the basal and stimulated ratios of areas are equal (see Figure 4.2). Therefore, we can easily compute the p-value of the test, by counting how often the observed statistic in absolute value, $|S^{obs}|$, exceeds the null distribution, $\hat{p} = \frac{1}{10^6}\sum_{p=1}^{10^6} \mathbb{1}(|S^{obs}| > |S^p|)$, where $\mathbb{1}(b)$ denotes the indicator variable which is 1 if condition $b$ is true, 0 otherwise. The observed p-value is 0.04, which suggests that the stimulant might have an impact on the relation between nuclear and cytoplasmic areas.

We also compute the nuclear and cellular areas after 1 and 2 hours from the first time the cell is observed, i.e. at times t = 31 and 61, expressed per units of

2 minutes as in the previous Chapter, to check whether the effect of the stimulant might change with time (Figures B.14 and B.15 in the Appendix). In these cases we only observe a subgroup of the above cells which are tracked in time, namely 25 under the basal condition and 40 under the stimulated one. We repeat the permutation test illustrated above and obtain p-values of 0.03 and 0.32, for t = 31 and 61 respectively, which suggests that the influence of the stimulation might decrease with time.

We also investigate whether the stimulation has an effect on the nuclear, cytoplasmic and cellular areas: we repeat the permutation test described above, for t = 1, 31 and 61, on the three areas alone. The p-values of the tests, given in Table 4.3, suggest the stimulation does not significantly influence the nuclear area but instead slightly shrinks the cytoplasmic one and hence, indirectly, has an effect on the cellular area and on the ratio. This effect seems to diminish after 2 hours, possibly indicating that the stimulation might not fully work any longer. The null distributions of the mean difference for these permutation tests are reported in the Appendix, in Figures B.16 and B.17.

| | t= 1 | t = 31 | t = 61 |
|---|---|---|---|
| Ratio | 0.039 | 0.029 | 0.321 |
| Nucleus | 0.056 | 0.357 | 0.461 |
| Cytoplasm | 0.001 | 0.033 | 0.233 |
| Cell | 0.002 | 0.042 | 0.247 |

**Table 4.3:** P-values of permutation tests on the mean of the ratio of areas and of nuclear, cytoplasmic and cellular areas, computed at time 1, 31 and 61 minutes.

Clearly this analysis only aims to get an approximate insight into this mechanism and is based on the assumption that the area of the 2 dimensional projection is a reasonable proxy for the volume.

Assuming that both nucleus and cell have a spherical shape, we can get an approximate insight into the ratio of volumes from their areas. If both nucleus and cell were exactly spherical, their volumes would be $\Omega_N = \frac{4}{3}\pi r_N^3$ and $\Omega_{Cell} = \frac{4}{3}\pi r_{Cell}^3$, respectively, where $r_N$ and $r_{Cell}$ represent the radii of nucleus and cell. Also, the area of the circles obtained as the 2 dimensional projection of nucleus and cell would be $A_N = \pi r_N^2$ and $A_{Cell} = \pi r_{Cell}^2$. Therefore, we could express the ratio between cytoplasmic and nuclear volumes, $c$, in terms of $A_N$ and $A_{Cell}$ as $c = \frac{\Omega_C}{\Omega_N} = \frac{\Omega_{Cell}}{\Omega_N} - 1 = \left(\frac{r_C}{r_N}\right)^3 - 1 = \left(\frac{A_{Cell}}{A_N}\right)^{3/2} - 1$.

These computations give rise only to a rough approximation of $c$ since, from the 2-dimensional images of our cells (see Figure 2.3), we notice that most cells do

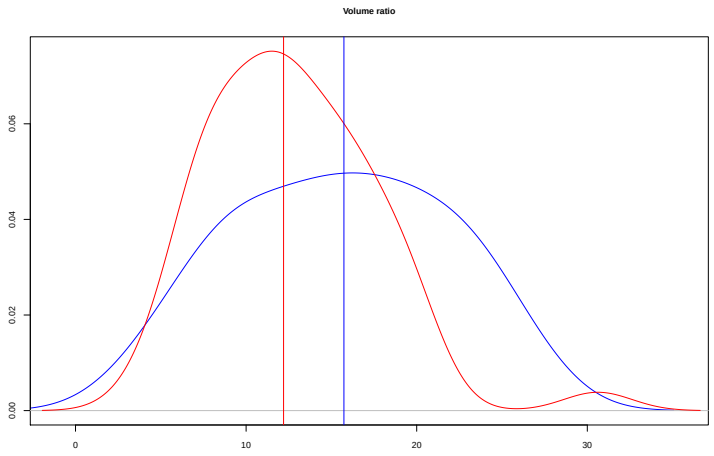


**Figure 4.3:** Estimated densities for the estimated volumes ratio $\hat{c}$, for the cells under the basal condition (in blue) and under the induction of a stimulant (in red). The vertical solid lines denote the mean values.

not have a circular representation, although this is usually approximately true for the nucleus. However, some information about possible values for $c$ can still be obtained to be fed into the priors. Considering cells at time t = 1, the estimated volume ratio varies, for the basal condition, between 3.8 and 26.4, with mean 15.5, and for the stimulated condition, between 4 and 30.6, with mean 12.7 (see Figure 4.3).

The mean of the logarithm of these values, 2.64 and 2.47 for the two conditions above, is used to formulate a distinct informative prior for the hypermean of $c^{(i)}$ for each condition, in the experimental data analysis, which will be presented in Section 4.6.

## 4.2 Exploratory analysis of import and export rates

Our collaborators believe that the time needed for Nrf2 to enter the nucleus in large quantities is longer than the time required for the same amount of Nrf2 to exit the nucleus. In other words, large exports from nucleus to cytoplasm are faster than equivalent imports into the nucleus.

We analyse the nuclear levels, in cells which exhibit clear oscillations, with no distinction between conditions, in order to understand whether there is evidence of a gradual increase, until a maximum is reached, and a quick drop after that. First, we identify 31 clear peaks and 15 troughs in our entire data, considering both conditions; we then compute the time needed to import and export roughly the same amount

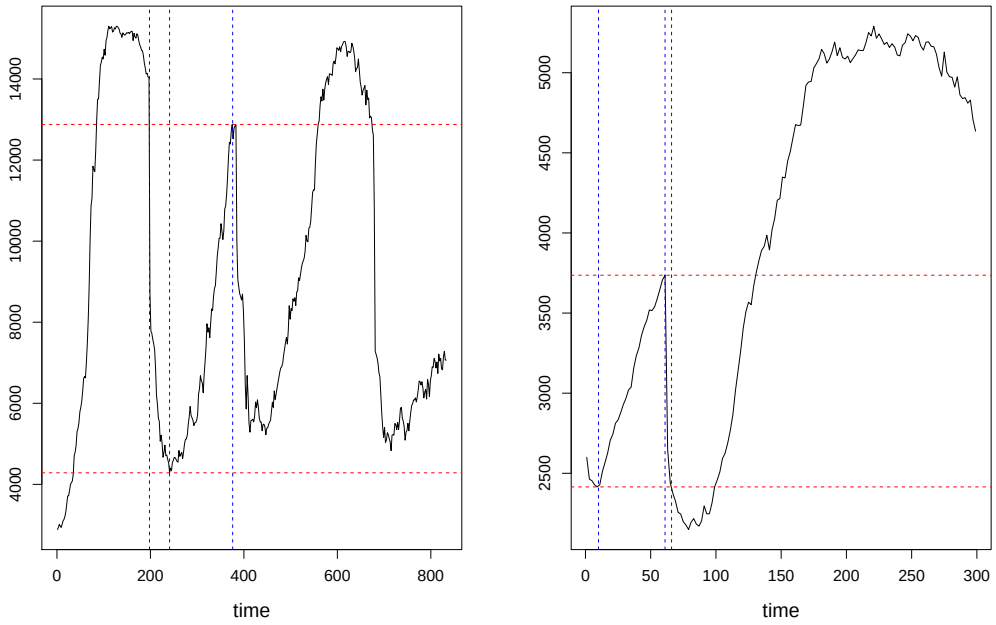


**Figure 4.4:** Example of the computation of nuclear import and export times on two cells centred, on the left, around a minimum and, on the right, around a maximum. The images show the observed nuclear concentration in two cells from our experimental data. The horizontal dashed lines represent the minimum and maximum levels of Nrf2 in a specific time interval. The vertical dashed lines indicate the times at which the nuclear concentration intercepts those values.

of Nrf2, that is, for almost identical initial and final levels (Figure 4.4). We define the former and latter times, expressed in minutes, as $t_I^{(i)}$ and $t_E^{(i)}$, for $i = 1, ..., 46$, where 46 is the total number of peaks and troughs we select in our collection.

We compute the ratio between import and export times (Figure 4.5) and, in agreement with our biologists' prior belief, we notice that, 45 times out of 46, the export happens in a smaller time interval than the import. This piece of information will also find confirmation in the dynamical behaviour of the estimated model which we study in Section 4.6.

In order to assess whether this difference is significant, we perform a permutation test (Fisher, 1935) on the mean time needed for import and export, which we call $E(t_I)$ and $E(t_E)$ respectively, and test the null hypothesis, $H_0 : E(t_I) = E(t_E)$, versus the alternative, $H_1 : E(t_I) \neq E(t_E)$. Similar to the previous Section, we consider the statistic $S = \bar{t}_I - \bar{t}_E$, where $\bar{t}_I = \sum_{i=1}^{46} t_I^{(i)}$ and $\bar{t}_E = \sum_{i=1}^{46} t_E^{(i)}$, in our

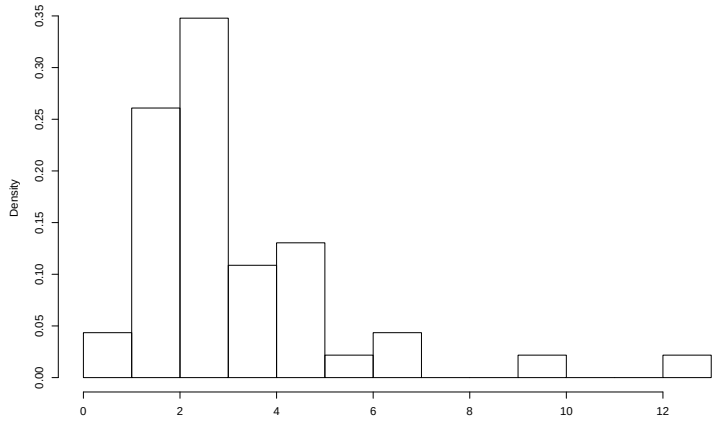


**Figure 4.5:** Histogram of the 46 observed ratios between import and export times, for the same amount of Nrf2, around a peak or a trough.

collection of 46 oscillations, $S = 49.6$ minutes. We sample the null distribution of $S$ by computing $10^6$ permutations of the import and export times and calculating, on each permutation, the mean difference between the first 46 values and second 46. The observed value of the statistic is at the far end of the right tail of the null distribution (Figure 4.6), with a p-value $< 10^{-5}$, indicating strong evidence that export and import of large quantities of Nrf2 require different times.

In particular, on average, the time needed for Nrf2 import into the nucleus is about 3.2 times bigger than for its export into the cytoplasm. We also notice that this average is 3.0 and 3.4 in the subsets of oscillations under the basal (29) and stimulated condition (17), respectively. This could suggest that the stimulant might slightly increase the difference between import and export times.

We will return to the issues introduced in this Section in the experimental data analysis described in Section 4.6.

## 4.3 Exploratory analysis of the measurement error

In the general case the measurement error covariance comprises three elements, as shown in (3.18). While it may normally seem plausible to assume a diagonal structure for $\Sigma_\epsilon$, which implies independent white noise error in distinct populations of molecules, it might not be necessarily appropriate in our context. In fact, before evaluating Nrf2 reporter light intensities in nucleus and cytoplasm, the border between the two compartments needs to be defined; this procedure is mostly manual

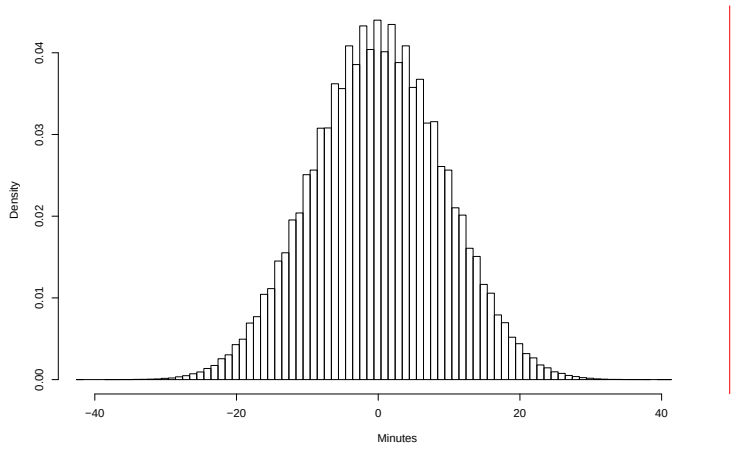


**Figure 4.6:** Null distribution of permutation test for the mean difference between import and export times. The vertical solid line on the far right represents the observed value of the statistic.

and is not exempt from error. Attributing a portion of the cell to the wrong compartment prompts a change in both nuclear and cytoplasmic light intensities. Therefore the measurement error variance matrix cannot, *a priori*, be assumed to be diagonal. Furthermore, the measurement error variances of the nuclear and cytoplasmic concentrations, although they refer to the same molecular population, are not necessarily equal. For instance, when identifying nuclear and cellular borders with the *CellTracking* software, the accuracy of the identification of the two boundaries may differ.

To keep notation clear, we drop the cell indicator $i$ from the following formulae, since most exploratory analyses are carried out independently on the three repeated measurements, unless specified.

Prior information about the three parameters of the covariance matrix can be obtained. Two sources of noise can be distinguished in $\epsilon_t$: one related to the light intensities' measurement process, which we call $\epsilon_t^L$, and one due to the identification of the nuclear and cellular boundaries in the *CellTracking* procedure, that we denote by $\epsilon_t^{CT}$. We believe that, in our data, the latter one represents the major source of measurement noise. By repeating this procedure twice on the same cell we get two realizations of the measurement process, $y^I = (y_1^I, ..., y_T^I)^T$ and $y^{II} = (y_1^{II}, ..., y_T^{II})^T$, both referring to the same original latent process $\tilde{x} = (\tilde{x}_1, ..., \tilde{x}_T)^T$, where

$$y_t^I = \tilde{x}_t + \epsilon_t^L + \epsilon_t^{CT^I} \quad \text{and} \quad y_t^{II} = \tilde{x}_t + \epsilon_t^L + \epsilon_t^{CT^{II}},$$

with $\epsilon_t^{CT^I}$ and $\epsilon_t^{CT^{II}}$ iid $\sim \mathcal{N}_2(0, \tilde{\Sigma}_\epsilon)$, $\forall t = 1, ..., T$, where $\tilde{\Sigma}_\epsilon$ represents the part of $\Sigma_\epsilon$ due to the cellular border identification, which we believe to dominate the noise covariance matrix (personal communication with Prof. Paul Thornalley and Dr Hiroshi Momiji).

For time reasons it was not feasible to implement this procedure on all observed cells, thus we collect repeated measurements on three randomly selected cells through which we infer parameters of

$$\tilde{\Sigma}_\epsilon = \begin{pmatrix} \tilde{\sigma}_N^2 & \tilde{\sigma}_{NC} \\ \tilde{\sigma}_{NC} & \tilde{\sigma}_C^2 \end{pmatrix}, \tag{4.1}$$

which, under the above assumptions, accurately approximates $\Sigma_\epsilon$, although under estimating it. The inferential results on $\tilde{\Sigma}_\epsilon$ obtained in this analysis will be used both to formulate an informative prior for $\Sigma_\epsilon$ parameters, and to define the structure of this matrix.

Figure 4.7 shows the observed nuclear and cytoplasmic concentrations, for the two measurements, in the 3 cells examined. In Figures 4.8 and 4.9 it is possible to observe the difference between the repeated measurements $y_t^I - y_t^{II} = \epsilon_t^{CT^I} - \epsilon_t^{CT^{II}}$ and the respective normal quantile-quantile plot; it is clear that the errors are autocorrelated in time.

There is a strong component of time dependence of the errors which is visible by observing the images showing the autocorrelation (ACF) of $\epsilon_t^{CT^I} - \epsilon_t^{CT^{II}}$ in the three cells in Figures B.18, B.19 and B.20. This indicates the inappropriateness of the independence assumption between errors at different time points. In fact, the borders identified are often copied from a frame to the next one and then adjusted to compensate for the cell movement. For instance, if at time $t$ the nuclear border is slightly misplaced and it causes an underestimate of the real nuclear concentration. Most likely at time $t + 1$ the nuclear border will be in a very near position and will give rise to a similar under estimate, and hence $\epsilon_{t+1}$ will be close to $\epsilon_t$. This prompts the need for a change in the nature of the white noise; a simple choice might consist of introducing an autoregressive (AR) model of order 1, which we denote by AR(1), $\epsilon_t = \phi\, \epsilon_{t-1} + z_t$, with $z_t \sim \mathcal{N}(0, \Sigma_z)$ and $\phi = diag(\phi_N, \phi_C)$. This modelling choice would imply the introduction of two new unknown parameters, $\phi_N$ and $\phi_C$ but it would also increase model realism. In fact, the partial autocorrelation (PACF), in Figures B.21, B.22 and B.23, drops after the first lag, as one would expect with an AR(1) model.

In simulations and experimental data analyses, we find that the AR(1) choice for the measurement error leads to a lack of practical identifiability of the error pa-

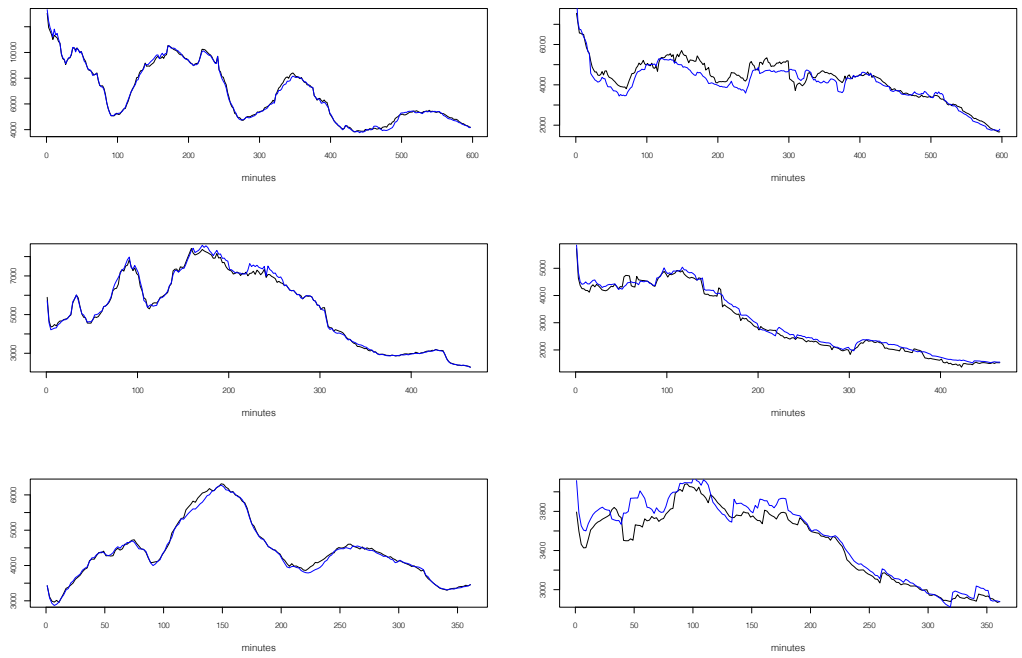


**Figure 4.7:** First (in black) and second (in blue) measurements, $y^I$ and $y^{II}$ respectively, of the nuclear (left) and cytoplasmic (right) concentrations in three cells, one on each row.

rameters, particularly for $\phi_N$ and $\phi_C$, which often tend to be overestimated. Therefore we use the normality assumption on the error which, although it ignores the time dependence between consecutive errors, represents the marginal distribution of $\epsilon_t$ in the AR model. We also carry out residual analyses to study the effect of our approximation on residuals. Furthermore, we perform a robustness analysis, in Section 4.4, by repeating the simulation studies shown in Section 3.7, this time simulating the measurement error from an AR(1) model, yet assuming a normal error in the estimation procedure. These simulation studies show inference on $\theta$ is robust with respect to the measurement model assumptions.

In order to understand whether we need to include the $\sigma_{NC}$ term in $\Sigma_\epsilon$, we perform a likelihood ratio test (LRT), on each one of the three cells independently, to test the following hypothesis system

$$\begin{cases} H0: & \tilde{\sigma}_{NC} = 0 \\ H1: & \tilde{\sigma}_{NC} \neq 0. \end{cases} \tag{4.2}$$

The likelihood of the non-constrained model, under $H_1$, is expressed as the product of normally distributed differences between the two realizations of the measurement

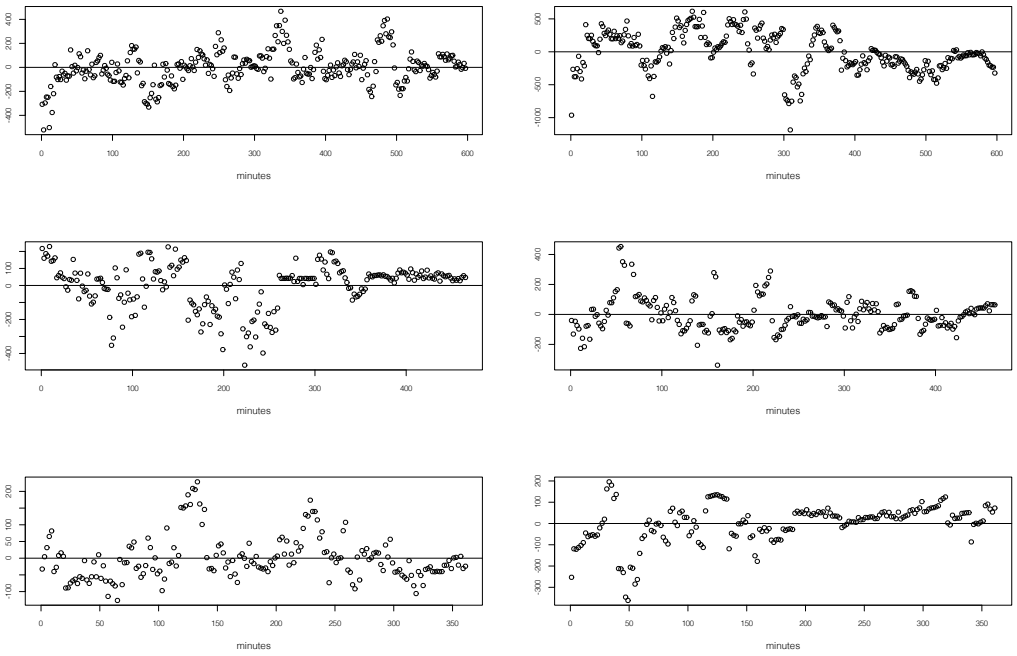


**Figure 4.8:** Differences between the repeated measurements, $y_t^I - y_t^{II} = \epsilon_t^{CT^I} - \epsilon_t^{CT^{II}}$, for the nuclear (left) and cytoplasmic (right) concentrations in the three cells, one on each row.

error, with zero mean and covariance matrix composed by three distinct elements,

$$L_{H_1}(\tilde{\sigma}_N, \tilde{\sigma}_C, \tilde{\sigma}_{NC}; y_t^I, y_t^{II}) = \prod_{t=1}^{T} \Phi\left( y_t^I - y_t^{II} | \mu = (0,0)^T, \Sigma = \begin{pmatrix} \tilde{\sigma}_N^2 & \tilde{\sigma}_{NC} \\ \tilde{\sigma}_{NC} & \tilde{\sigma}_C^2 \end{pmatrix} \right).$$

Similarly, the likelihood under the null hypothesis, $H_0$, is

$$L_{H_0}(\tilde{\sigma}_N, \tilde{\sigma}_C; y_t^I, y_t^{II}) = \prod_{t=1}^{T} \Phi\left( y_t^I - y_t^{II} | \mu = (0,0)^T, \Sigma = \begin{pmatrix} \tilde{\sigma}_N^2 & 0 \\ 0 & \tilde{\sigma}_C^2 \end{pmatrix} \right).$$

The LRT, for each cell, is defined as

$$\psi = 2\left(log(L_{H_1}(\hat{\tilde{\sigma}}_N^{(H_1)}, \hat{\tilde{\sigma}}_C^{(H_1)}, \hat{\tilde{\sigma}}_{NC}^{(H_1)}; y_t^I, y_t^{II})) - log(L_{H_0}(\hat{\tilde{\sigma}}_N^{(H_0)}, \hat{\tilde{\sigma}}_C^{(H_0)}; y_t^I, y_t^{II}))\right),$$

where $\hat{\tilde{\sigma}}_N^{(H_1)}$, $\hat{\tilde{\sigma}}_C^{(H_1)}$ and $\hat{\tilde{\sigma}}_{NC}^{(H_1)}$ represent the maximum likelihood estimates (MLEs) for $\tilde{\sigma}_N$, $\tilde{\sigma}_C$ and $\tilde{\sigma}_{NC}$, under $H_1$, while $\hat{\tilde{\sigma}}_N^{(H_0)}$ and $\hat{\tilde{\sigma}}_C^{(H_0)}$ are the MLEs for $\tilde{\sigma}_N$ and $\tilde{\sigma}_C$, under $H_0$. The LRT, under $H_0$, is asymptotically distributed as a $\psi \sim \chi_1^2$ (Pace and Salvan, 1997). Two of the three tests are not significant, at the 5% level, with

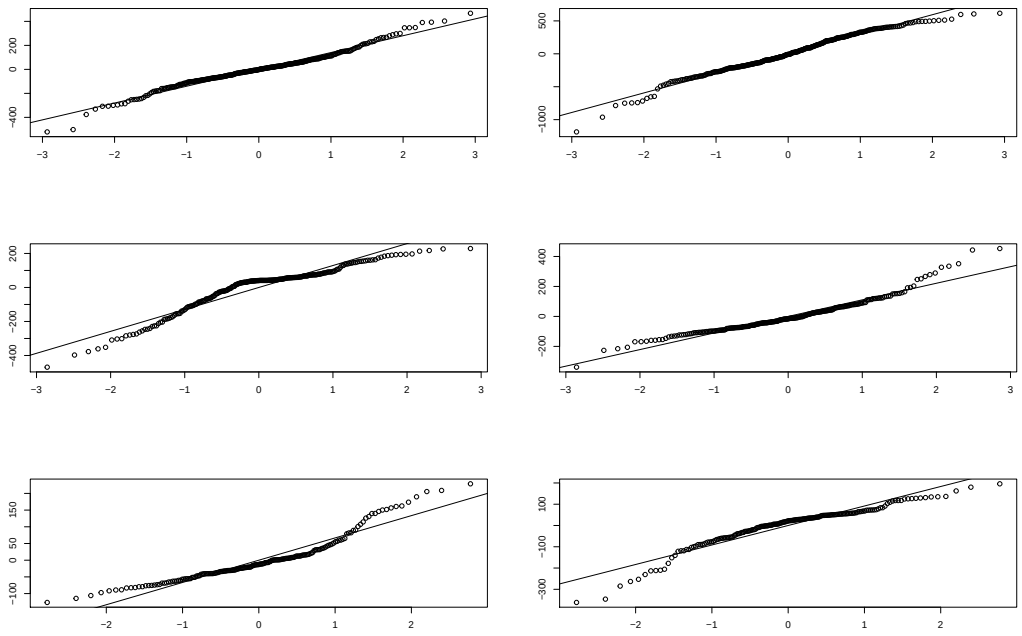

**Figure 4.9:** Quantile-quantile plot of the difference between repeated measurements, $y_t^I - y_t^{II} = \epsilon_t^{CT^I} - \epsilon_t^{CT^{II}}$, for the nuclear (left) and cytoplasmic (right) concentrations in the three cells, one on each row.

p-values of 1.000, 0.099 and 0.015. Although the significant p-value in the third cell, from the ACF plots in Figures B.18-B.20, we notice that, in all three cells, there seem to be no strong linear dependence between the errors in nucleus and cytoplasm.

We also test, via the LRT, whether it may plausible to assume a non-hierarchical $\Sigma_\epsilon$, constant for all cells, via the hypothesis system:

$$\begin{cases} H0: & \tilde{\sigma}_N^{(i)} = \tilde{\sigma}_N \text{ and } \tilde{\sigma}_C^{(i)} = \tilde{\sigma}_C, \text{ for } i = 1,2,3 \\ H1: & \text{distinct } \tilde{\sigma}_N^{(i)} \text{ and } \tilde{\sigma}_C^{(i)} \text{ across cells, } i = 1,2,3, \end{cases} \tag{4.3}$$

where the $(i)$ superscript indicates that the parameter refers to the $i$-th cell.

First, we need to define the following hierarchical quantities: $D = (D^{(1)}, D^{(2)}, D^{(3)})$, where $D^{(i)} = (D_1^{(i)}, ..., D_{T_i}^{(i)})^T$, with $D_t^{(i)} = (y_t^{I^{(i)}} - y_t^{II^{(i)}})$ representing the difference between the two measurements of the same process at time $t$, $\tilde{x}_t^{(i)}$, for $i = 1,2,3$; furthermore we introduce $\theta_\sigma = (\theta_\sigma^{(1)}, \theta_\sigma^{(2)}, \theta_\sigma^{(3)})$, where $\theta_\sigma^{(i)} = (\tilde{\sigma}_N^{(i)}, \tilde{\sigma}_C^{(i)})^T$, for $i = 1,2,3$.

In this case, assuming independence between the thee cells, the full likelihood

under $H_1$ becomes the product of the likelihoods of the three cells

$$L_{H_1}(\theta_\sigma; D) = \prod_{i=1}^{3} \prod_{t=1}^{T_i} \Phi \left( D_t^{(i)} \,|\, \mu = (0,0)^T, \Sigma = \begin{pmatrix} \tilde{\sigma}_N^{2(i)} & 0 \\ 0 & \tilde{\sigma}_C^{2(i)} \end{pmatrix} \right). \tag{4.4}$$

While, under $H_0$ the likelihood is obtained as above, yet assuming the covariance matrix parameters are the same for all cells,

$$L_{H_0}(\tilde{\sigma}_N, \tilde{\sigma}_C; D) = \prod_{i=1}^{3} \prod_{t=1}^{T_i} \Phi \left( D_t^{(i)} \,|\, \mu = (0,0)^T, \Sigma = \begin{pmatrix} \tilde{\sigma}_N^2 & 0 \\ 0 & \tilde{\sigma}_C^2 \end{pmatrix} \right).$$

The LRT in this case, under $H_0$, is asymptotically distributed as a $\chi_4^2$, where the four degrees of freedom are given by the difference between the parameters of the alternative and null models, 6 and 2 respectively. In the null model indeed we assume common nuclear and cytoplasmic variances across cells, therefore only 2 parameters are required for the covariance matrices of all cells. In the alternative instead, two distinct parameters are required in each of the three cell considered here, for a total of 6 parameters.

The value of the test is highly significant, with p-value $< 10^{-3}$, indicating a strong cell-to-cell variability of the measurement error variance and, hence, highlighting the importance of treating these parameters hierarchically, allowing them to vary between cells.

Finally, we test, again in the three cells separately, whether there is a significant difference between the variances of the nuclear and cytoplasmic errors via the following hypothesis system:

$$\begin{cases} H0: & \tilde{\sigma}_N = \tilde{\sigma}_C \\ H1: & \tilde{\sigma}_N \neq \tilde{\sigma}_C. \end{cases} \tag{4.5}$$

The LRT for each cell, under $H_0$, asymptotically follows a $\chi_1^2$ distribution. As expected, al tests are highly significant at the 5% level, with p-values $< 10^{-3}$, 0.020 and $< 10^{-3}$.

The results of these tests allow us to formulate the measurement error covariance matrix for the $i$-th cell as a diagonal matrix, composed of two distinct variance parameters for the nuclear and cytoplasmic measurement errors,

$$\Sigma_\epsilon^{(i)} = \begin{pmatrix} \sigma_N^{2(i)} & 0 \\ 0 & \sigma_C^{2(i)} \end{pmatrix}. \tag{4.6}$$

We also find that we need to treat the measurement error parameters hierarchically, as they significantly differ from cell to cell. The preliminary results obtained in this Section will be confirmed by the residual analysis of the experimental data analyses which we describe in Section 4.7, where we will see that residuals are not found to be correlated between nucleus and cytoplasm.

### Hierarchical analysis on repeated measurements

We perform a Bayesian hierarchical analysis on the repeated measurement data in $D$, with the full likelihood being $L_{H_1}$ defined in (4.4), to infer the nuclear and cytoplasmic standard deviations of the error, in $\theta_\sigma$.

The error parameters in each cell are assumed to be distributed, *a priori*, $\sigma_N^{(i)} \sim log\mathcal{N}\left(\mu_N, \frac{1}{\tau_N}\right)$ and $\sigma_C^{(i)} \sim log\mathcal{N}\left(\mu_C, \frac{1}{\tau_C}\right)$, for $i = 1, 2, 3$. As in the full analysis, we use a conjugate normal-gamma prior for the hyper-parameters, with $\mu_j|\tau_j \sim \mathcal{N}\left(0, \frac{10^4}{\tau_j}\right)$, for $j = N, C$, and $\tau_j \sim \mathcal{G}(0.001, 0.001)$, for $j = N, C$, where the former is the normal r.v. with zero mean and variance $\frac{10^4}{\tau_j}$ and the latter is the gamma distribution with mean 1 and variance $10^3$. We implement a simple Bayesian hierarchical analysis, where posterior densities of the parameters are estimated via MCMC methods as follows.

**Initialisation**: First we initialise parameters in $\theta_\sigma$ and hyperparameters $\mu_N$, $\mu_C$, $\tau_N$ and $\tau_C$ to arbitrary positive real values.

**Update**: At each iteration of the MCMC, $r = 1, ..., R$, with $R$ being the total number of iterations, we alternately sample from the conditional distributions of the hyperparameters and of the hierarchical parameters.

hyperparameters: Due to the conjugacy of the hyperprior distribution, proceeding as in Section 3.6, the hyperparameters are sampled from a Gibbs step where, for $j = N, C$,

$$\mu_j|\tau_j \sim \mathcal{N}\left(\frac{\tau_j \sum_{i=1}^{3} log(\sigma_j^{(i)})}{\tau_j/10^4 + 3\,\tau_j}, (\tau_j/10^4 + 3\,\tau_j)^{-1}\right),$$

and

$$\tau_j|\mu_j \sim Gamma\left(0.001 + 3/2,\, 0.001 + 1/2 \sum_{i=1}^{3} \left(log(\sigma_j^{(i)}) - \mu_j\right)^2\right).$$

Hierarchical parameters: For each $i = 1, ..., 3$, $(\sigma_N^{(i)}, \sigma_C^{(i)})$ is jointly sampled from a Metropolis step, from a normal proposal with adaptive covariance. We call $(\sigma_{N_*}^{(i)}, \sigma_{C_*}^{(i)})$ the newly proposed values, while $(\sigma_N^{(i)}, \sigma_C^{(i)})$ represent the current iteration values.

The acceptance rate for $(\sigma_{N_*}^{(i)}, \sigma_{C_*}^{(i)})$ is $min(1, A_\sigma)$, where

$$A_\sigma = \frac{\prod_{j\in\{N,C\}} \Phi(log(\sigma_{j_*}^{(i)})|\mu_j, \frac{1}{\tau_j})\, L_\sigma(D^{(i)}; \sigma_{N_*}^{(i)}, \sigma_{C_*}^{(i)})}{\prod_{j\in\{N,C\}} \Phi(log(\sigma_j^{(i)})|\mu_j, \frac{1}{\tau_j})\, L_\sigma(D^{(i)}; \sigma_N^{(i)}, \sigma_C^{(i)})},$$

$$\text{with } L_\sigma(D^{(i)}; \sigma_N, \sigma_C) = \prod_{t=1}^{T_i} \Phi\left(D_t^{(i)} \,|\, \mu = (0,0)^T, \Sigma = \begin{pmatrix} \sigma_N^2 & 0 \\ 0 & \sigma_C^2 \end{pmatrix}\right).$$

We apply this algorithm on our repeated measurements on three cells. Figure B.24 shows the histograms for the inferred posterior densities of the hypermeans, $\mu_N$ and $\mu_C$, and hyperprecisions, $\tau_N$ and $\tau_C$; we can also observe the posterior chains for the four hyperparameters in Figure 4.10. We compute the posterior modes of $\mu_N$ and $\mu_C$ to be 4.41 and 4.52, respectively. The posterior modes are computed in  (R Core Team, 2016) via the package *MCMCglmm* (Hadfield, 2016).

These results are used to formulate an informative prior, for the hypermeans of $\sigma_N$ and $\sigma_C$. In particular, for both conditions, we set hyperparameters for $\mu_N$ and $\mu_C$, as shown in Section 3.5, as $\mu_{j_0} = 4.41$ and $\lambda_{j_0} = 0.1$ for the nuclear standard deviation, and $\mu_{j_0} = 4.52$ and $\lambda_{j_0} = 0.1$, for the cytoplasmic one. However, given the tiny sample of 3 cells used in this analysis, we chose not to formulate an informative prior for the precision hyperparameters in order not to inform our model *a priori* about how measurement error standard deviations vary between cells. The informative hyperprior for $\mu_N$ and $\mu_C$ will be used in the experimental data analysis, in Section 4.6.

## 4.4 Robustness simulation study

In order to study what effect autocorrelated measurement error might have on inference when erroneously assuming independent error terms across time, we perform again inference on the 3 simulation datasets analysed in Section 3.7; this time we substitute the uncorrelated measurement error component with a correlated one.

In particular, for the $i$-th cell of each simulation, $i = 1, ..., N$, we simulate the measurement error $\epsilon^{(i)}$ from an autoregressive models of order 1:

$$\epsilon_t^{(i)} = \phi^{(i)}\, \epsilon_{t-1}^{(i)} + z_t^{(i)}, t = 1, ..., T_i$$

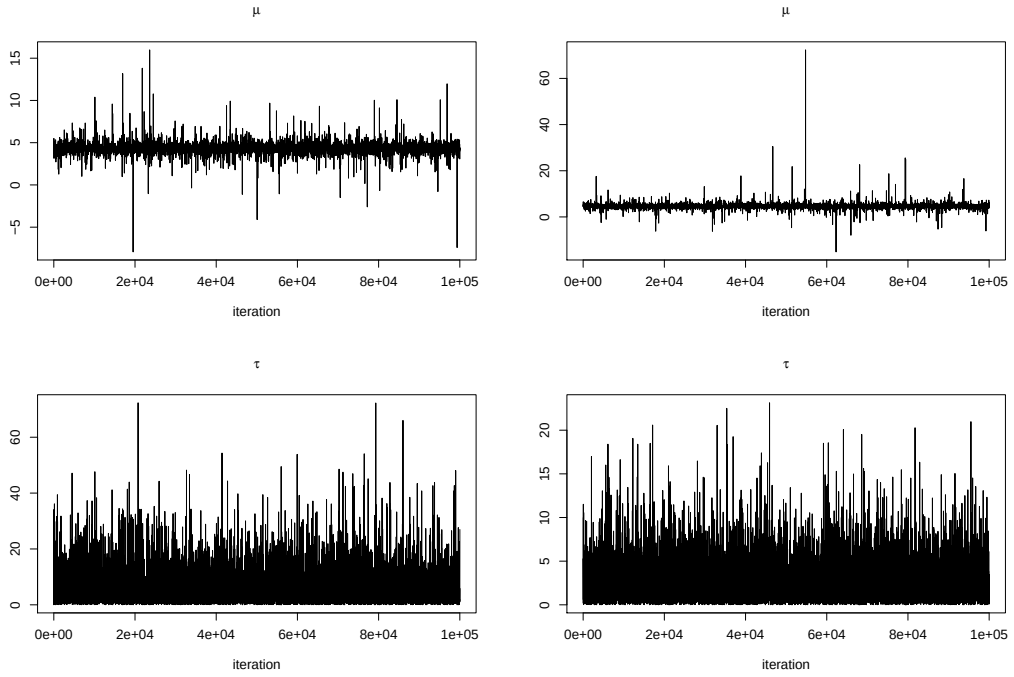


**Figure 4.10:** MCMC posterior chains, *burn-in* excluded, for the hypermean, $\mu_N$ and $\mu_C$ (top row), and hyperprecision,$\tau_N$ and $\tau_C$ (bottom row), parameters.

with $\phi^{(i)} = diag(\phi_N^{(i)}, \phi_C^{(i)})$ and $z_t^{(i)} \sim \mathcal{N}(0, \Sigma_z^{(i)})$, where $\Sigma_z^{(i)} = diag(\sigma_{z_N}^{2(i)}, \sigma_{z_C}^{2(i)})$. Nuclear and cytoplasmic error components are, as in the previous simulation study, sampled separately.

We then analyse these simulated datasets with the same methodology as in Section 3.7, assuming a normal error, and compute coverages of the 0.99 and 0.95 HPD CIs (Table 4.4). We find that coverages are very close to the nominal ones and similar to those reported in the Section 3.7.

Coverages are also satisfactory for the hyperparameters: all hyperprecisions and 23 out of 24 hypermeans fall in the respective 0.99 and 0.95 level CIs.

| Level | $k_d^{(i)}$ | $k_a^{(i)}$ | $K_a^{(i)}$ | $\mu_\tau^{(i)}$ | $\gamma^{(i)}$ | $\delta^{(i)}$ | $c^{(i)}$ | $\kappa_N^{(i)}$ | Tot |
|---|---|---|---|---|---|---|---|---|---|
| 99 | 99.0 | 100.0 | 98.1 | 98.1 | 100.0 | 100.0 | 99.0 | 99.0 | 99.2 |
| 95 | 91.4 | 100.0 | 98.1 | 92.4 | 97.1 | 100.0 | 88.6 | 99.0 | 95.8 |

**Table 4.4:** Empirical coverages, in percentage, of 0.99 and 0.95 level HPD credible intervals, based on 105 hierarchical parameters from 3 simulations, each with 35 simulated processed. The last column indicates the average coverage of parameters ($k_d^{(i)}$, $k_a^{(i)}$, $K_a^{(i)}$, $\mu_\tau^{(i)}$, $\gamma^{(i)}$, $\delta^{(i)}$, $c^{(i)}$, $\kappa_N^{(i)}$).

We also compute coverages for the marginal standard deviation of the error. In fact, in an AR(1) model, we can compute the marginal variance of each component (Mills, 1991) as $Var(\epsilon_{N_t}^{(i)}) = \tilde{\sigma}_N^{2(i)} = \dfrac{\sigma_{z_N}^{2^{(i)}}}{1-\phi_N^{2^{(i)}}}$ and $Var(\epsilon_{C_t}^{(i)}) = \tilde{\sigma}_C^{2(i)} = \dfrac{\sigma_{z_C}^{2^{(i)}}}{1-\phi_C^{2^{(i)}}}$. Table 4.5 reports coverages of the CIs for $\tilde{\sigma}_N^{(i)}$ and $\tilde{\sigma}_C^{(i)}$: as expected, coverages are lower than the nominal levels, as in the previous simulation, due to the misspecification of the error, but not excessively low.

| Level | $\tilde{\sigma}_N^{(i)}$ | $\tilde{\sigma}_C^{(i)}$ |
|---|---|---|
| 99 | 99.0 | 92.4 |
| 95 | 92.4 | 85.7 |

**Table 4.5:** Empirical coverages, in percentage, for the marginal standard deviation of the measurement error parameters, of 0.99 and 0.95 level HPD credible intervals, based on 105 hierarchical parameters from 3 simulations, each with 35 simulated processed.

Overall, inference on the 3 simulated datasets is successful and inference on the parameters of interest does not seem to be affected by the misspecification of the measurement equation. In Section 4.6 we apply the methodology developed to both experimental datasets, in order to infer the model parameters.

## 4.5 Population levels

In this Section we introduce some prior information, from the literature as well as the knowledge of our collaborators, about the total molecular population of Nrf2 reporter; this will allow us to constrain the parameter space of the proportionality constants $\kappa_N^{(i)}$ and $\kappa_C^{(i)}$.

Biggin (2011) presents a survey, from the literature, of several reliable estimates of transcription factors (TFs) in humans and animals, and indicates that most animal TFs are expressed at 10,000-300,000 molecules per nucleus. More specifically for Nrf2 protein, our biological collaborators believe the population of our TF to be between 5,000 and 50,000 molecules (personal communication with Prof. Paul Thornalley). In fact, Nrf2 is a low copy regulatory TF, which is present with fewer molecules than high abundance housekeeping proteins.

Xue *et al.* (2015b) estimate that, on the same data we analyse, the reporter Nrf2 only induces a minor increase in the total Nrf2 pool of 4-7%. This increase refers to the entire population of cells considered; however, the mean overall transfection was only 40% (personal communication with two of the authors, Prof. Paul Thornalley and Dr Hiroshi Momiji). By overall transfection, we refer to the transient

transfection process, described in Section 2.2, where we insert into cells an engineered version of the DNA, that is able to transcribe reporter mRNA which then translated into the fluorescent reporter protein that we observe. This process is not always successful and, sometimes, we fail to observe the reporter Nrf2. In particular, in our data, the overall tranfection is successful in about 40% of the cells. Therefore, if we only consider the fraction of cells where transfection was successful, which are the ones we analyse in this study, Nrf2 reporter induces an increase of up to 10-17.5% of the total molecular population.

We can use these pieces of information to gain an understanding of the range of possible values for the total number of molecules of Nrf2 reporter in a cell, i.e. $X_t^{N(i)} + X_t^{C(i)}$ at time $t$ for $i$-th cell. While there is more uncertainty regarding a plausible upper bound for this interval, we can more easily formulate a conservative lower bound. First, we consider 5,000 as the lower bound for the original Nrf2 population in an entire cell, which is the minimum of the two lower bounds described above. This corresponds to 200-350 molecules of Nrf2 reporter in a cell, assuming the 4-7% proportion, and to 500-875, for the more realistic 10-17.5% estimate.

Furthermore, the light intensities in the available cells, as visible in Figure 2.3, are very homogeneous and one cannot distinguish single molecules by eye; this indicates that each cell has many molecules of reporter Nrf2, probably in the order of hundreds, to create a smooth light intensity when stimulated by a laser.

Say we assume a general lower bound, which we call $\tilde{min}_X$, and limit the cellular population of Nrf2 reporter to have at least $\tilde{min}_X$ proteins in each cell we analyse and where transfection process was successful. This translates into having at least $\tilde{min}_X$ molecules, on average over the observational time, for each latent process $X_t^{(i)} = (X_t^{N(i)}, X_t^{C(i)})^T$, as $E_t(X_t^{N(i)} + X_t^{C(i)}) > \tilde{min}_X$.

We can re-formulate this constraint in terms of observed processes, $Y_t^{(i)}$, by inverting the measurement equation in (3.21) and exploit the fact that the error has zero mean, as

$$E_t(X_t^{N(i)} + X_t^{C(i)}) \simeq E_t\left(\frac{Y_t^{N(i)}}{\kappa_N^{(i)}} + \frac{Y_t^{C(i)}}{\kappa_C^{(i)}}\right) > \tilde{min}_X, \tag{4.7}$$

where $\kappa_C^{(i)}$ is obtained as $\dfrac{\kappa_N^{(i)}}{c^{(i)}} = \dfrac{\kappa^{(i)}}{\Omega_C^{(i)}} = \kappa_C^{(i)}$.

To implement this constraint, we simply limit parameters $\kappa_N^{(i)}$ and $c^{(i)}$ to respect (4.7).

In order to choose an optimal $\tilde{min}_X$, we study, via the analysis shown below,

some plausible values for it: 100, 200, 500 and 1,000. We compare how the inferred parameters, under each constraint, are able to mimic the observed data; in particular we consider the autocorrelation (ACF) function as a proxy for the oscillatory behaviour.

### Study of the constraint

We use the methodology described in Chapter 3 to infer the posterior distributions of the parameters from the experimental data, under both conditions. We repeat the full analysis on both our experimental data sets four times, each time using a different constraint in equation (4.7): 100, 200, 500 and 1,000. In this way, we obtain, for every experimental condition, four complete posterior densities for each parameter. All four analyses produce similar results with respect to the difference between the two conditions.

For each constraint and condition, we select 100 parameter values from the MCMC posterior chains of every cell, excluding *burn-in*; values are chosen to be equally spaced along the chains in order to diminish their dependence and obtain almost independent draws from their posterior distributions.

From each of the 100 selected parameter vectors, we simulate, via the DA and measurement equation (3.21) as in Section 3.7, a process for $Y^{(i)}$. Therefore, for every constraint, we obtain 100 simulated processes per cell, for a total of 3,500 processes for the basal condition and 3,600 for the stimulated one.

In every cell, we compare the estimated autocorrelation (ACFs) function of each simulated process with the ACFs of the original data; in particular, we compute the sum of the absolute differences, for nucleus and cytoplasm, between the ACFs, of lag 1 to 60 minutes, of original and simulated data. Therefore, for each cell and simulation, we obtain a number estimating how well the simulated data mimics the oscillatory pattern of the experimental data. We then average these quantities, over the 100 simulations and over the cells, and obtain, for every constraint, one value for the basal and one for the stimulated condition, indicating how closely, overall, the simulations emulate the ACFs of the observed data, where the ACF is taken as a proxy for the oscillatory behaviour.

Figure 4.11 shows the sum of absolute differences for the various constraints: both conditions exhibit similar patterns with a clear minimum at 200.

In the next Section, we show inferential results, obtained from the available experimental data assuming, via constraint (4.7), that at least $\tilde{min}_X = 200$ molecules of reporter protein are present in each cell.

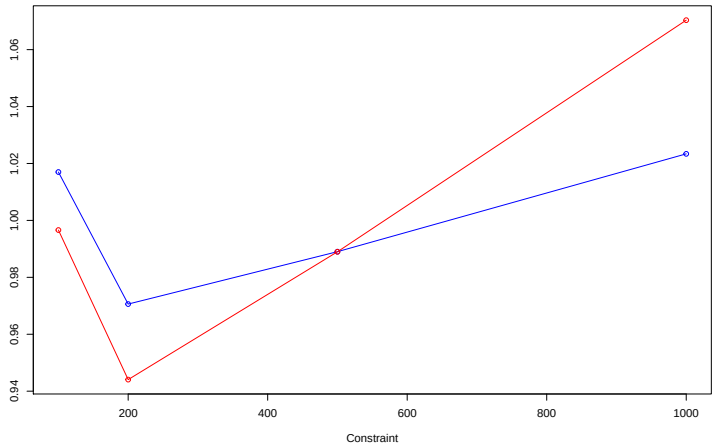


**Figure 4.11:** Sum of absolute differences of autocorrelations for the basal (blue line) and stimulated (red curve) conditions, for constraint (4.7) equal to 100, 200, 500 and 1,000 (horizontal axis).

## 4.6 Data analysis

We apply the methodology described in Chapter 3 to infer parameters' posterior distributions from the available data, independently, on both basal and stimulated conditions. All rates and parameters are expressed per minute. The average acceptance rate for each point of the latent process is about 68% and 72 % for the basal and stimulated conditions, respectively. The acceptance rates for the five blocks of hierarchical parameters, $(k_d^{(i)}, \mu_\tau^{(i)})$, $(k_a^{(i)}, K_a^{(i)})$, $(\delta^{(i)}, \gamma^{(i)})$, $(c^{(i)}, \kappa_N^{(i)})$ and $(\sigma_N^{(i)}, \sigma_C^{(i)})$, are 33, 25, 42, 26 and 7%, for the basal condition, and 37, 22, 48, 26 and 8% for the stimulated one.

After discarding the initial part of each chain, as a *burn-in*, we run the MCMC for further $4 \times 10^5$ iterations, which we use to approximate the parameters' posterior distributions.

Figure 4.12 shows, for both conditions, the traceplot for the logarithm of the full DA function in (3.24), $log(L_{DA}(\theta; \tilde{\mathbf{X}}))$. Figures B.25 and B.26 in the Appendix show the traceplots for the hypermeans of the kinetic paramaters, after discarding the *burn-in*. The traceplots for the single cell DA, $log(L_{DA}(\theta^{(i)}; \tilde{x}^{(i)}))$, are shown in Figures B.27-B.28.

Figure B.29 displays the posterior densities of all hierarchical parameters, while Figures B.30 and B.31 report the posterior densities of the hypermean and hyperprecision parameters.

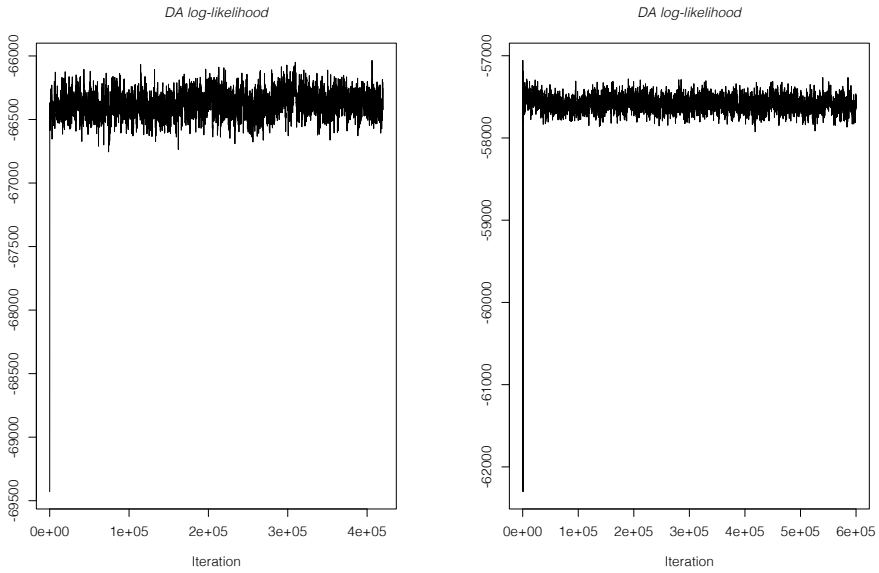


**Figure 4.12:** Chains for the full $L_{DA}$ in (3.24), for the experimental data analysis under the basal (left) and stimulated conditions (right).

Of particular interest is Figure 4.13, which shows the posterior densities of the exponential of the hypermeans which, in a log-normal distribution, correspond to the medians of hierarchical parameters; Table 4.6 reports the 0.95 level HPD CI for the same parameters.

The differences between the two conditions are not striking; this is reasonable, considering that stimulated cells are induced at a level ($2\mu$M) which is compatible with bioactive compounds that can be found in fruit and vegetables.

The most evident difference between the basal (in blue) and stimulated (in red) conditions is observed for parameters $k_d$ and $k_a$ in the top row; in particular, the latter has higher rates than the former. There also seems to be an interesting difference between the exponential of the hypermeans in the forth row, for parameters $c^{(i)}$ and $\kappa_N^{(i)}$. In particular, the ratio between cytoplasmic and nuclear volumes, $c^{(i)}$, appears to be, on average, slightly higher for the basal condition, which is coherent with our exploratory finding in Section 4.1.

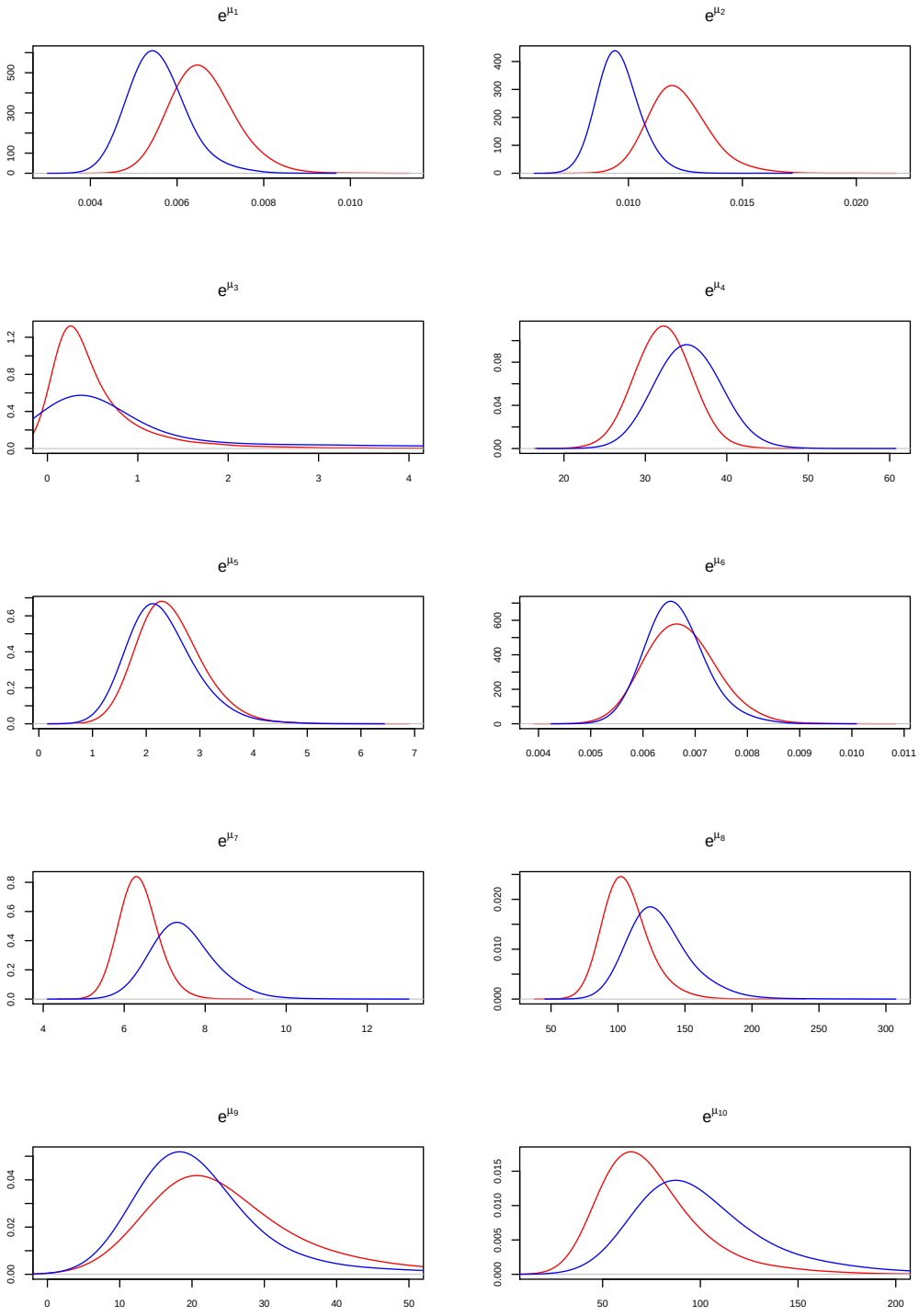


**Figure 4.13:** Posterior densities for the exponential of the hypermean, $e^{\mu_j}$, corresponding to the median of hierarchical parameters, under the basal (blue) and stimulated (red) conditions.

| Parameter | Basal | | Stimulated | |
|---|---|---|---|---|
| | LB | UB | LB | UB |
| $k_d^{(i)} \times 10^3$ | 3.76 | 5.71 | 4.75 | 6.91 |
| $k_a^{(i)} \times 10^3$ | 7.34 | 10.41 | 8.97 | 12.79 |
| $K_a^{(i)}$ | 0.11 | 3.38 | 0.11 | 1.83 |
| $\mu_\tau^{(i)}$ | 24.83 | 39.11 | 24.32 | 36.38 |
| $\gamma^{(i)}$ | 1.21 | 2.81 | 1.43 | 3.12 |
| $\delta^{(i)} \times 10^3$ | 5.14 | 7.00 | 4.99 | 7.05 |
| $c^{(i)}$ | 5.55 | 7.58 | 5.24 | 6.57 |
| $\kappa_N^{(i)}$ | 69.67 | 121.29 | 61.68 | 101.68 |
| $\sigma_N^{(i)}$ | 2.85 | 17.47 | 3.82 | 17.33 |
| $\sigma_C^{(i)}$ | 28.00 | 61.87 | 22.98 | 47.44 |

**Table 4.6:** HPD CI of level 0.95 for the exponential of the hypermean, $e^{\mu_j}$, for both conditions; LB and UB denote the lower and upper bounds of the CI, respectively.

A straightforward comparison between import and export times cannot be carried out from Figure 4.13; in fact the former follows a linear rate, while the latter is modelled by a non-linear Michaelis-Menten (MM) term. In order to compare the two, we compute the average of the MM term over the observational time which, for cell $i$, is defined as $m^{(i)} = \left\langle \frac{k_a^{(i)} X_t^{N(i)}}{K_a^{(i)} + X_t^{N(i)}} \right\rangle$, where $\langle\, . \,\rangle$ denotes the average operator, across the $T_i$ observations of our process. In practice, we approximately compute this quantity from observations $Y^{(i)}$, taking advantage of the fact that $\kappa X^{(i)}$ dominates $\epsilon^{(i)}$, which also has zero mean, as

$$m^{(i)} = \left\langle \frac{k_a^{(i)} X_t^{N(i)}}{K_a^{(i)} + X_t^{N(i)}} \right\rangle = \left\langle \frac{k_a^{(i)} \tilde{X}_t^{N(i)}}{\kappa_N^{(i)}\, K_a^{(i)} + \tilde{X}_t^{N(i)}} \right\rangle \simeq \left\langle \frac{k_a^{(i)} Y_t^{N(i)}}{\kappa_N^{(i)}\, K_a^{(i)} + Y_t^{N(i)}} \right\rangle = \hat{m}^{(i)}.$$

In order to compare the linear import $k_d^{(i)}$ with $\hat{m}^{(i)}$, we average the hierarchical values across the cells and obtain, under both conditions, the posterior densities for their means, $\frac{1}{N}\sum_{i=1}^{N} k_d^{(i)}$ and $\frac{1}{N}\sum_{i=1}^{N} \hat{m}^{(i)}$. Figure 4.14 reports the posterior densities for these quantities in both conditions.

We notice how the stimulation (in red) induces faster import and export with respect to the basal condition (in blue), justifying the faster movements observed in the stimulated experimental data. This finding will also be confirmed in the stability analysis, in Chapter 5, where it is shown that a higher fraction of stimulated cells is associated to noise-induced oscillatory patterns, as compared to the cells under the basal condition.

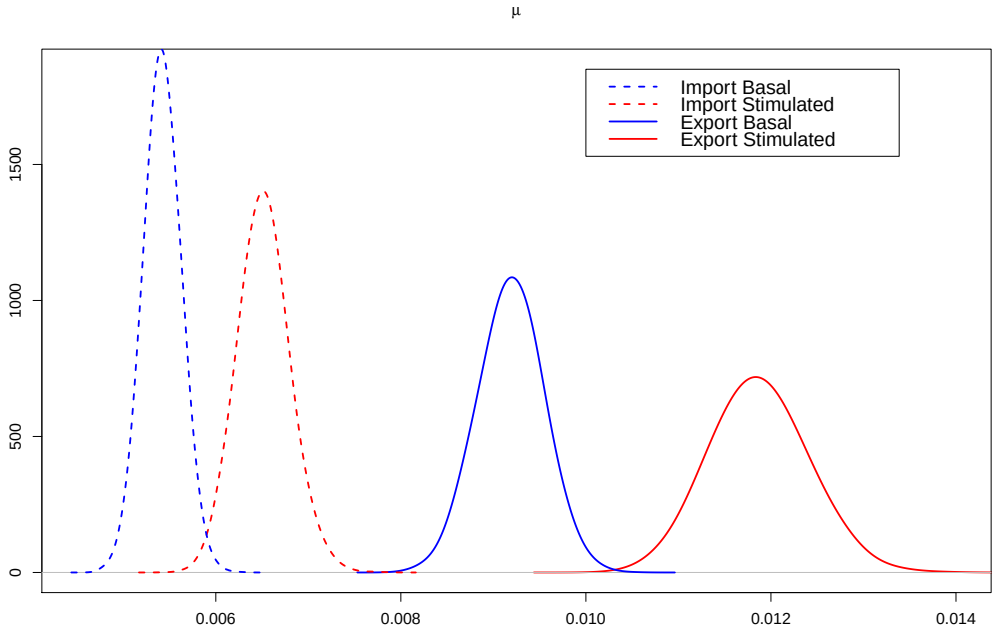


**Figure 4.14:** Posterior densities for the mean, over the hierarchical parameters, of the nuclear import rate, $\frac{1}{N}\sum_{i=1}^{N} k_d^{(i)}$ (dashed line), and Michaelis-Menten export rate, $\frac{1}{N}\sum_{i=1}^{N} \hat{m}^{(i)}$ (solid line), under the basal (blue) and stimulated conditions (red).

Figure 4.14 also indicates that, in both conditions, the export clearly happens at a faster rate than the import, confirming the preliminary findings of Section 4.2. Finally, we notice that there is no clear evidence that the stimulation might increase this difference between export and import times, as hinted by the analysis in Section 4.2.

We also consider the ratio between synthesis and degradation rates, which represents the total population of Nrf2 reporter proteins at stationarity. To see this, we consider the ordinary differential equation (ODE) for the total population of Nrf2 in a cell which at time $t$, in the deterministic context, we call $Z_t^{cell} = Z_t^N + Z_t^C$, where the $Z_t^N$ and $Z_t^C$, as defined in Section 2.5, are the nuclear and cytoplasmic deterministic counterparts of the stochastic populations $X_t^N$ and $X_t^C$. The ODE for $Z_t^{cell}$ reduces to:

$$\dot{Z}_t^{cell} = \gamma - \delta\, Z_t^{cell},$$

which has stationary solution in $x_{cell}^* = \dfrac{\gamma}{\delta}$. Therefore, the ratio between synthesis and degradation indicates the population of Nrf2 reporter the cell tends to. Figure

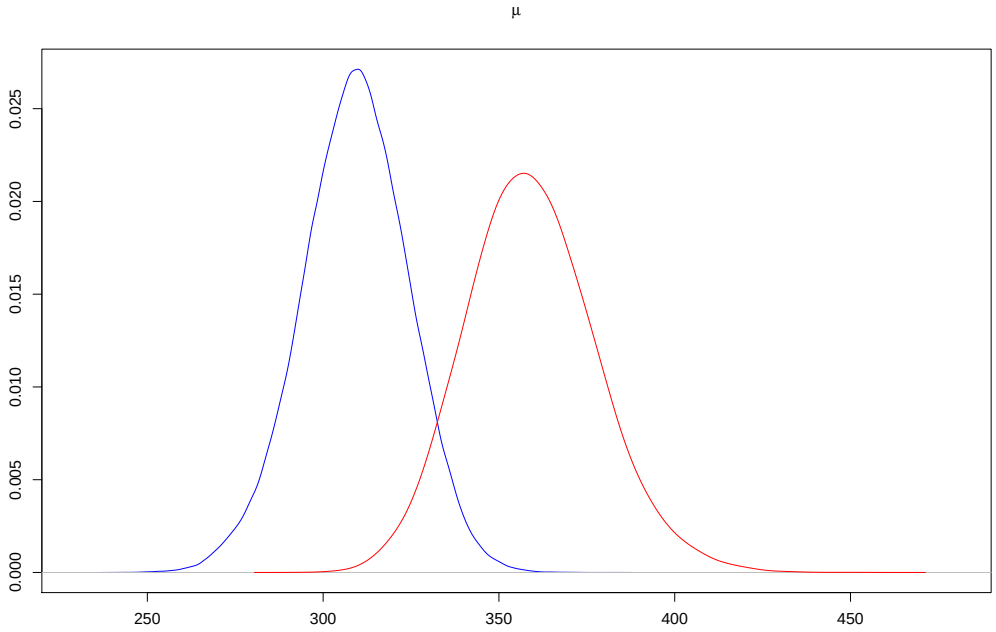


**Figure 4.15:** Posterior densities for the ratio between synthesis and degradation: mean of hierarchical parameters, under the basal (blue) and stimulated conditions (red).

4.15 shows the posterior densities for the average, across cells, of the ratio between synthesis and degradation, $\frac{1}{N}\sum_{i=1}^{N}\frac{\gamma^{(i)}}{\delta^{(i)}}$. The densities for both conditions are well above 200, the minimum number of molecules we assumed.

Results from this Section indicate that cells in the two conditions approximately follow similar mechanisms. Stimulation does not appear to strongly affect cellular processes such as synthesis and degradation; on the contrary it may have an impact on the nuclear and cytoplasmic volumes. The most interesting finding is that stimulation seems to induce faster movements between nucleus and cytoplasm, which translate in slightly more oscillatory patterns. Furthermore, in both conditions, we also show confirmatory evidence that export happens at a faster rate than import, as already suggested in the analysis in Section 4.2.

In Chapter 5 we will again make use of the posterior chains obtained in this Section to perform a stability analysis to further investigate the nature of Nrf2 oscillations.

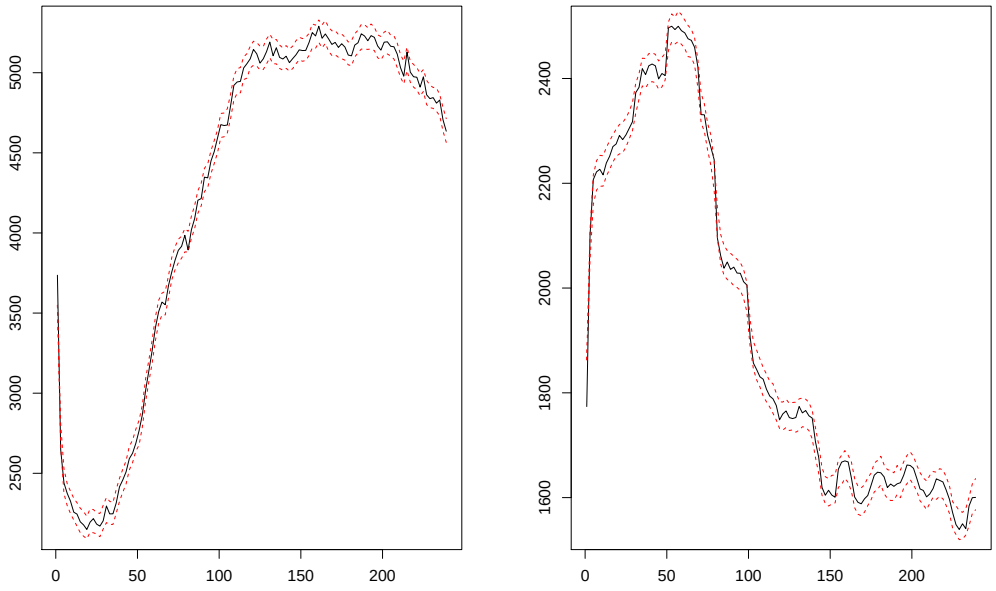


**Figure 4.16:** Observed Nrf2 reporter levels $Y^{(i)}$ (black solid line) in a nucleus (left) and cytoplasm (right), and 0.95 CI of the reconstructed latent process $\tilde{X}^{(i)}$ (red dashed lines) for a single cell.

## 4.7 Residual analysis

We analyse the residuals of the model obtained, for each observation $Y_t^{(i)} = (Y_t^{N(i)}, Y_t^{C(i)})^T$, as the difference between the observation and the reconstructed latent process, $\tilde{X}_t^{(i)} = (\tilde{X}_t^{N(i)}, \tilde{X}_t^{C(i)})^T$, which at the $k$-th iteration of the MCMC algorithm, we call $\tilde{x}_t^{(i)^k} = (\tilde{x}_t^{N(i)^k}, \tilde{x}_t^{C(i)^k})^T$, where superscripts $N$ and $C$ refer to the nuclear and cytoplasmic components respectively. In particular, at the $k$-th iteration of the MCMC, for $k = 1, ..., K$, we define the column vectors for the nuclear and cytoplasmic residuals for the $i$-th cell, with $i = 1, ..., N$, as $r_N^{(i)^k} = (r_{N_1}^{(i)^k}, ..., r_{N_{T_i}}^{(i)^k})^T$ and $r_C^{(i)^k} = (r_{C_1}^{(i)^k}, ..., r_{C_{T_i}}^{(i)^k})^T$, where

$$r_{N_t}^{(i)^k} = Y_t^{N(i)} - \tilde{x}_t^{N(i)^k} \text{ and } r_{C_t}^{(i)^k} = Y_t^{C(i)} - \tilde{x}_t^{C(i)^k}.$$

All nuclear and cytoplasmic residuals for the $i$-th cell are embedded in the $T_i \times K$ matrices $r_N^{(i)} = (r_N^{(i)^1}, ..., r_N^{(i)^K})$ and $r_C^{(i)} = (r_C^{(i)^1}, ..., r_C^{(i)^K})$, respectively. For computer memory and storage reasons, we apply a thinning factor of $10^3$ to the residuals, i.e.

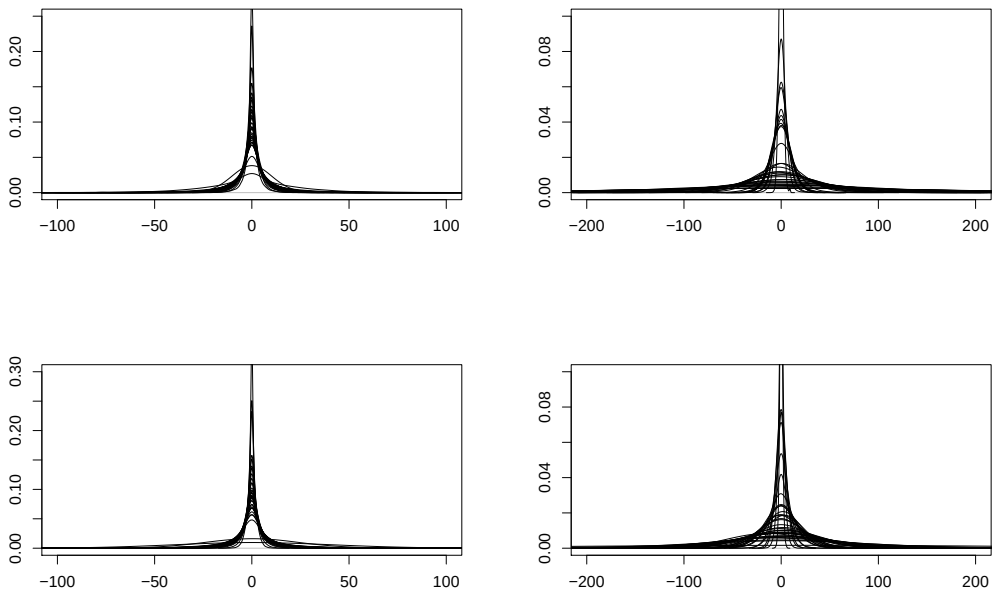


**Figure 4.17:** Densities for nuclear (left) and cytoplasmic (right) residuals observed in the 35 cells under the basal (top row) and stimulated (bottom row) conditions.

only 1 residual process is stored every $10^3$ iterations; indexes $k = 1, ..., K$ refer to the thinned chains.

Therefore, for every cell $i = 1, ..., N$, each iteration of the MCMC has its own nuclear and cytoplasmic residual vectors, $r_N^{(i)^k}$ and $r_C^{(i)^k}$. Figure 4.16 shows, for a particular cell, the observed process, $Y^{(i)}$, and the pointwise 0.95 HPD CI for the latent process, $\tilde{X}^{(i)}$, which clearly is in the neighbourhood of the observed one. The CI is obtained, for each $\tilde{X}_t^{N(i)}$ and $\tilde{X}_t^{C(i)}$, from the respective chains $(\tilde{x}_t^{N(i)^1}, ..., \tilde{x}_t^{N(i)^K})$ and $(\tilde{x}_t^{C(i)^1}, ..., \tilde{x}_t^{C(i)^K})$.

Figures B.32-B.43 in the Appendix report, for every cell $i$, the histograms of residuals for all observations of each cell, $r_N^{(i)}$ and $r_C^{(i)}$, under both conditions: they all appear to be symmetric and centred in 0. Figure 4.17 summarizes the information from the histograms by displaying the overlapping densities, one for each cell $i$, of all residuals in nucleus, $r_N^{(i)}$, and cytoplasm, $r_C^{(i)}$, for the basal and stimulated conditions.

We also check for autocorrelation in the residuals: in each cell $i$, and for every MCMC iteration $k$, we compute the autocorrelation of nuclear and cytoplasmic residuals of lag $j = 1, ..., 20$, which we define as $A_{N_k}^{(i)^j} = acf_j(r_N^{(i)^k})$ and $A_{C_k}^{(i)^j} = acf_j(r_C^{(i)^k})$, respectively, where $acf_j(.)$ measures the empirical autocorre-

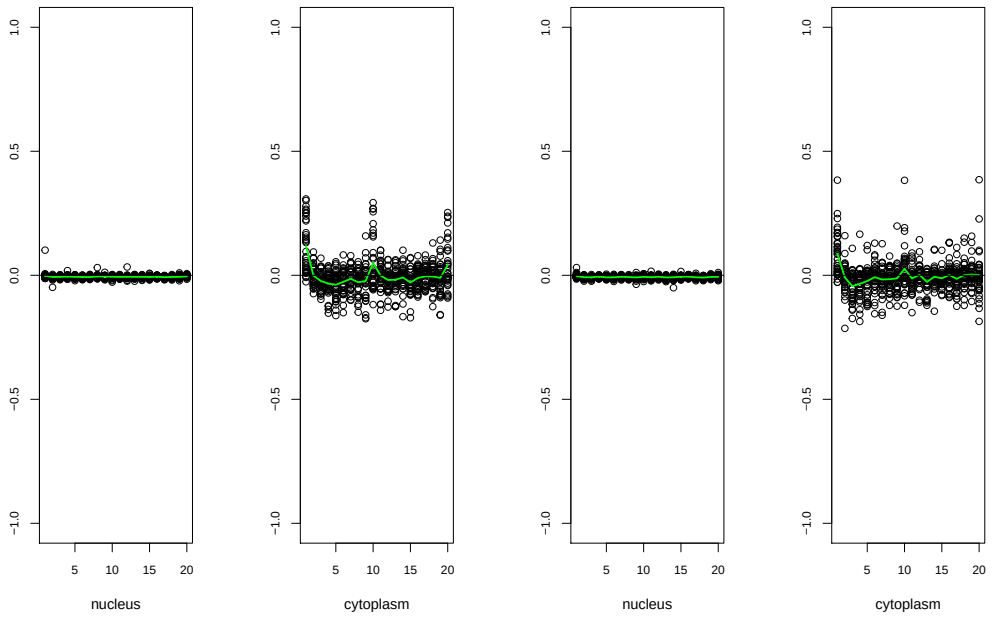


**Figure 4.18:** Autocorrelation of residuals of lag from 1 to 20 which, given observations are available every 2 minutes, corresponds to a maximum time lag of 40 minutes. Each point represents the average, over the MCMC iterations, autocorrelation of residuals in a cell; the green line indicates the average over cells. The left two plots refer to nuclear (first left) and cytoplasmic (second left) autocorrelation under the basal condition. The third and fourth panels give analogous results for the stimulated condition.

lation function of lag $j$. We average these values across the MCMC iterations, to obtain a unique ACF for nucleus and cytoplasm of each cell $i$, which we call $\bar{A}_N^{(i)^j} = \frac{1}{K}\sum_{k=1}^{K} A_{N_k}^{(i)^j}$ and $\bar{A}_C^{(i)^j} = \frac{1}{K}\sum_{k=1}^{K} A_{C_k}^{(i)^j}$. Figure 4.18 shows the average autocorrelation across the MCMC iterations, of lag $j = 1, ..., 20$, of the nuclear and cytoplasmic residuals, $\bar{A}_N^{(i)^j}$ and $\bar{A}_C^{(i)^j}$; every dot refers to a cell. The points appear to be homogeneously spread around 0; this is particularly true if we consider the average across cells, $\frac{1}{N}\sum_{i=1}^{N} \bar{A}_N^{(i)^j}$ and $\frac{1}{N}\sum_{i=1}^{N} \bar{A}_C^{(i)^j}$, denoted by the green line. There only seems to be a slightly positive autocorrelation of lag 1 for the cytoplasmic residual component, although 3 cases, for both basal ($3/35 = 8.6\%$) and stimulated ($3/36 = 8.3\%$) conditions, have negative autocorrelations.

Finally, we study the linear dependence between nuclear and cytoplasmic residuals. Again, for each cell $i$ and MCMC iteration $k$, we compute the correla-

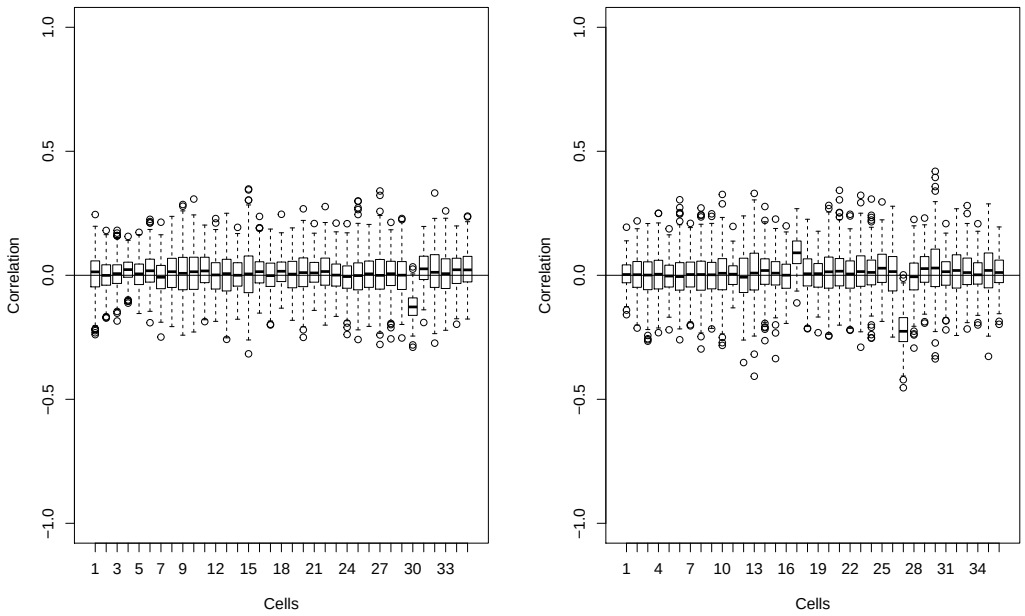


**Figure 4.19:** Boxplots for the covariance between nuclear and cytoplasmic errors; each boxplot corresponds to a cell in the basal (left) and stimulated (right) conditions.

tion between the nuclear and cytoplasmic residual vectors, which we call $C_k^{(i)} = cor(r_N^{(i)^k}, r_C^{(i)^k})$, where $cor(a, b)$ indicates the empirical correlation between $a$ and $b$. For each cell, we obtain $K$ correlations $C_1^{(i)}, ..., C_K^{(i)}$: Figure 4.19 shows the boxplots obtained from these elements, one for each cell, under both basal (left image) and stimulated (right plot) conditions. The image clearly indicates that there is no evidence of any linear dependence between nuclear and cytoplasmic residuals.

Overall, residual analyses show the measurement error assumptions are fulfilled and coherent with the data: residuals are symmetric around 0 and are neither autocorrelated in time nor correlated between compartments.

## 4.8 Summary

This Chapter has focused on the analysis of the available experimental data from the Nrf2 system. In the first part we presented exploratory analyses which allow us to gain an initial insight into our data. In particular, by studying the ratio between nuclear and cytoplasmic areas, we are able to formulate an informative prior for the hypermean of the volumes ratio. Following an analysis on repeated measurements,

we also obtain an informative prior on the hypermeans of the measurement error parameters. Furthermore, we show that nuclear and cytoplasmic estimated errors do not appear to be correlated, although they might be autocorrelated in time. In order to investigate the robustness of the model to a misspecification of the error, we perform a further simulation study where the error is sampled from an AR(1) model. Finally, we find evidence that the Nrf2 nuclear export is faster than its import; the same piece of information is then confirmed by the posterior results obtained when fitting the model to data.

We then apply the methodology shown in the previous Chapter to our experimental data. We do not find large differences in the posterior densities of parameters between the conditions; however, as expected, we find that stimulated cells seem to have faster import and export rates, which would explain the more dynamic and oscillatory pattern observed in the experimental data when stimulated. We also check that residuals are symmetric around zero and are neither correlated between nucleus and cytoplasm nor autocorrelated in time.

In the next Chapter, we will further exploit the parameters' posterior chains we obtained. By means of a stability analysis of the system around its equilibrium, we will obtain further insight into the mechanism driving the oscillations we observe in some of our experimental data.

# Chapter 5

# Stability analysis of the Nrf2 system

I saw the angel in the marble and carved until I set him free.

*Michelangelo Buonarroti*

In this Chapter we study the stability of the Nrf2 system, around its stationary solution. We initially analyse the model in a deterministic context and study its linear approximation near equilibrium. We then investigate what happens when perturbing the stationary solution with a source of white noise. Finally, we draw some conclusions about the nature of the system.

## 5.1 DDE at equilibrium

We start by studying the behaviour of the deterministic system, defined by the following delay differential equations (DDEs),

$$\dot{Z}_t^N = k_d\, Z_t^C - E_\tau(Z_{t-\tau}^N)\,\frac{k_a\, Z_t^N}{K_a + Z_t^N} - \delta Z_t^N \qquad \text{[ Nuc Nrf2 ]} \qquad (5.1)$$

$$\dot{Z}_t^C = E_\tau(Z_{t-\tau}^N)\,\frac{k_a\, Z_t^N}{K_a + Z_t^N} - k_d\, Z_t^C + \gamma - \delta Z_t^C, \qquad \text{[ Cyt Nrf2 ]} \qquad (5.2)$$

where, for simplicity, we drop the cell indicator $^{(i)}$ and, as explained in Section 2.5, $Z_t^N$ and $Z_t^C$ represent the deterministic counterparts of the stochastic populations $X_t^N$ and $X_t^N$.

We select, from the posterior chains of each cell, *burn-in* excluded, 100 parameter vectors. The chosen samples are equally spaced along the posterior chain, so as to reduce their dependence and obtain almost independent draws. Thus, in total we select 3,500 parameter vectors under the basal condition and 3,600 under the stimulated one. For each cell and parameter vector, we then use the Euler algorithm (Wilkinson, 2012) to numerically integrate each one of the 7,100 DDEs. We find that all DDEs quickly converge towards an equilibrium and none of them converges to a stable periodic oscillation. In the Euler integration, the delay component, $E_\tau(Z^N_{t-\tau})$, which is an integral over the delay distribution, is approximated by a discrete summation over the past values of the process, up to lag $\tau^{max} = 60$ minutes, as in (2.4).

## 5.2 Linear stability analysis

We are interested in understanding what happens when we introduce small perturbations around the equilibrium. In order to do so, we study the nature of the eigenvalues of the linearised system around its equilibrium (Hirsch *et al.*, 2013; Momiji and Monk, 2009; Strogatz, 2014; Luenberger, 1979). First, we write the DDE at its stationary solution, which we define as $x^* = (x^*_N, x^*_C)$, as

$$0 = k_d\, x^*_C - x^*_N\, \frac{k_a\, x^*_N}{K_a + x^*_N} - \delta x^*_N \qquad [\text{ Nuc Nrf2 }] \tag{5.3}$$

$$0 = x^*_N\, \frac{k_a\, x^*_N}{K_a + x^*_N} - k_d\, x^*_C + \gamma - \delta x^*_C, \qquad [\text{ Cyt Nrf2 }] \tag{5.4}$$

where 0 on the left side of either equation is due to the fact that, at equilibrium, the derivative, with respect to time, of the deterministic process is zero.

Recall that, as explained in Section 4.6, the stationary solutions for nucleus and cytoplasm are connected, because the equilibrium for the entire cell, $Z^{cell}_t$, is the ratio between synthesis and degradation parameters, $x^*_{cell} = x^*_N + x^*_C = \dfrac{\gamma}{\delta}$.

We consider equation (5.1) around its equilibrium and rewrite $Z^N_t$ as its stationary values plus a small perturbation, $Z^N_t = x^*_N + w_t$. We also assume the total cellular level is stable near equilibrium $Z^{cell}_t \simeq \frac{\gamma}{\delta}$; hence, in (5.1), we can rewrite $Z^C_t$ as $Z^C_t = Z^{cell}_t - Z^N_t \simeq x^*_{cell} - Z^N_t = \dfrac{\gamma}{\delta} - x^*_N - w_t = x^*_C - w_t$. In other words, we separate the whole Nrf2 dynamics into two parts: the behaviour of the total cellular amount of Nrf2 and the nuclear-cytoplasmic translocations; the latter is the component we are interested in investigating here. By assuming that the total Nrf2 is constant at equilibrium, we focus on the DDE for $Z^N_t$ alone and, hence, reduce

the problem from 2 dimensions to a single one.

In order to linearise (5.1), we linearise the Michaelis-Menten term via a first order Taylor expansion around the stationary solution:

$$\frac{k_a\, Z_t^N}{K_a + Z_t^N} \simeq \frac{k_a\, x_N^*}{K_a + x_N^*} + w_t\, \frac{k_a\, K_a}{(K_a + x_N^*)^2},$$

where we neglect second order terms as we have assumed to be near equilibrium, where $w_t$ is small.

Hence, we rewrite equation (5.1), in its linearised form, as

$$\dot{Z}_t^N = k_d\, Z_t^C - E_\tau(Z_{t-\mu_\tau}^N)\, \frac{k_a\, Z_t^N}{K_a + Z_t^N} - \delta Z_t^N \tag{5.5}$$

$$\simeq k_d\, (x_C^* - w_t) - (x_N^* + E_\tau(w_{t-\tau}))\, \left( \frac{k_a\, x_N^*}{K_a + x_N^*} + w_t\, \frac{k_a\, K_a}{(K_a + x_N^*)^2} \right) - \delta\, (x_N^* + w_t) \tag{5.6}$$

$$\simeq a^* + w_t \left( -k_d - \delta - x_N^*\, \frac{k_a\, K_a}{(K_a + x_N^*)^2} \right) - E_\tau(w_{t-\tau})\, \frac{k_a\, x_N^*}{K_a + x_N^*} \tag{5.7}$$

$$\simeq w_t \left( -k_d - \delta - x_N^*\, \frac{k_a\, K_a}{(K_a + x_N^*)^2} \right) - E_\tau(w_{t-\tau})\, \frac{k_a\, x_N^*}{K_a + x_N^*} \tag{5.8}$$

where, going from 5.6 to 5.7, we drop the second order term $w_{t-\tau}\, w_t\, \frac{k_a\, K_a}{(K_a + x_N^*)^2}$ and, in 5.7, $a^* = \left( k_d\, x_C^* - x_N^*\, \frac{k_a\, x_N^*}{K_a + x_N^*} - \delta\, x_N^* \right)$, which equals 0 as $a^*$ corresponds to the DDE for $Z_t^N$, computed at equilibrium, in (5.3).

The DDE for $Z_t^N$, around its stationary value, is replaced by the DDE for $w_t$. Since $Z_t^N = x_N^* + w_t$, and $x_N^*$ is constant, it follows that $\dot{Z}_t^N = \dot{w}_t$ (Strogatz, 2014; Luenberger, 1979). After linearisation, $\dot{w}_t$ is approximately linear and, hence, can be expressed as $\dot{w}_t \simeq \lambda\, w_t$, where $\lambda$ is defined as the eigenvalue solution of the linearised system we are interested in studying (Strogatz, 2014; Luenberger, 1979). The trivial solution of $\dot{w}_t$ at time $t$, is $w_t = w_0\, e^{\lambda t}$, where $w_0$ represents its starting value at time 0 (Strogatz, 2014; Luenberger, 1979).

We replace, in equation (5.8), $\dot{Z}_t^N$ with $\dot{w}_t$ and $w_t$ with $w_0\, e^{\lambda t}$. After the $w_0$ and $e^{\lambda t}$ terms cancel out on the two sides of the equation, we can solve for $\lambda$:

$$\lambda\, w_0\, e^{\lambda t} \simeq w_0\, e^{\lambda t} \left( -k_d - \delta - x_N^*\, \frac{k_a\, K_a}{(K_a + x_N^*)^2} \right) - w_0\, E_\tau(e^{\lambda(t-\tau)})\, \frac{k_a\, x_N^*}{K_a + x_N^*} \tag{5.9}$$

$$\lambda \simeq \left( -k_d - \delta - x_N^*\, \frac{k_a\, K_a}{(K_a + x_N^*)^2} \right) - E_\tau(e^{-\lambda\tau})\, \frac{k_a\, x_N^*}{K_a + x_N^*} \tag{5.10}$$

$$\lambda \simeq a + b\, E_\tau(e^{-\lambda\tau}), \tag{5.11}$$

where $a = -k_d - \delta - x_N^* \dfrac{k_a\, K_a}{(K_a + x_N^*)^2}$, $b = -\dfrac{k_a\, x_N^*}{K_a + x_N^*}$ and the integral in $E_\tau(e^{-\lambda\tau}) = \displaystyle\int_0^{\tau_{max}} e^{-\lambda\tau} f_\tau(\tau) d\tau$ is solved numerically via the function *integrate* provided by the  software (R Core Team, 2016).

Alternatively, we could proceed in a similar manner with the second DDE for $Z_t^C$, i.e. assuming $Z_t^{cell}$ constant and studying the intracellular mechanisms via the 1-dimensional system for $Z_t^C$, which mirrors the one for $Z_t^N$. Clearly, this approach would lead to the same eigenvalue equation. We do not report calculations for brevity, as they are specular to the ones presented above.

The value of $\lambda$ determines the linear stability of the steady state solution. In particular, the real part of the eigenvalue determines if, when perturbing the system, this perturbation decays, for a negative real part, or grows, for a positive one (Strogatz, 2014; Luenberger, 1979). This result becomes intuitive if considering the DE for the perturbation $\dot{w}_t \simeq \lambda\, w_t$. However, negative complex eigenvalues, i.e. with non-zero imaginary part, correspond to systems whose fixed points are stable spirals (Strogatz, 2014). We are particularly interested in these cases because such systems, although deterministically converging to a stable solution, are likely to produce transient oscillatory behaviours when induced by noise. Therefore, in a stochastic context, when constantly perturbed by noise, these systems oscillate around the equilibrium solution. Such cases are often referred to as "noise-induced oscillators", since the oscillations are not inherent in the system itself, and are only observable in a noisy context. Noise-induced oscillations are a well known phenomenon, present in many biological mechanisms (Steuer, 2004; Ko *et al.*, 2010; Balanov *et al.*, 2004).

## 5.3 Eigenvalue study

We numerically solve the eigenvalue equation (5.11), for each of the 7,100 sampled parameter vectors and corresponding stationary solutions. All our solutions have negative real part as, in the equation for $\lambda$, both $a$ and $b$ are negative, implying that, when perturbing the system, the noise decays in time. In addition, we find complex eigenvalues in 49% of cases, under the basal condition, and in 67%, under the stimulated one. This finding suggests that, at least in a good portion of cells, the Nrf2 system behaves like a noise-induced oscillator. In other words, although the deterministic system does not exhibit sustained oscillations, the stochasticity introduced by the finite population often makes the system oscillate around its solution.

Moreover, the fact that the fraction of complex eigenvalues is higher in stimulated cells, again confirms that, when cells are induced, they are more likely to oscillate.

| Model | Basal | Stimulated | Mode Basal | Mode Stimulated |
|---|---|---|---|---|
| Original model | 53.0 | 66.5 | 48.6 | 58.3 |
| Constant $\tau = \mu_\tau$ | 52.3 | 65.9 | 45.7 | 58.3 |
| $2 \times \sigma_\tau$ | 55.0 | 68.6 | 54.3 | 61.1 |
| $4 \times \sigma_\tau$ | 61.1 | 74.7 | 60.0 | 66.7 |
| Linear export | 52.5 | 66.3 | 48.6 | 58.3 |
| Constant $\tau = 0$ | 0.0 | 0.0 | 0.0 | 0.0 |
| Constant $\tau = 60$ | 97.8 | 98.9 | 100.0 | 100.0 |
| Constant $\tau = 120$ | 100.0 | 100.0 | 100.0 | 100.0 |

**Table 5.1:** Number of complex solutions, in percentage, for the eigenvalue analysis. "Original model" refers to the model used for inference on the experimental data. In "Constant $\tau = \mu_\tau$" the delay is assumed to be a constant equal its mean parameter, which varies according to the parameter vector used. Rows "$2 \times \sigma_\tau$" and "$4 \times \sigma_\tau$" correspond to the model with standard deviation of the delay distribution 2 and 4 times bigger than the original one (3 minutes), i.e. $\sigma_\tau$ is 6 and 12 minutes respectively. In "Linear export" the MM term is replaced by a constant, $\theta_{MM}$. The bottom three lines, "Constant $\tau = 0$, 60 or 120", indicate that the delay is constant and equal to 0, 60 or 120 minutes.

Next, we investigate the effect of the delay and the Michaelis-Menten (MM) term. In particular, we repeat the stability analysis by varying the standard deviation for the delay distribution, where we use a standard deviation, $\sigma_\tau$, 2 and 4 times bigger than the original one, i.e. $\sigma_\tau$ is set equal to 6 and 12 minutes, whilst 3 minutes is the original value used in the data analysis. We also consider a constant delay instead of a distributed one, which corresponds to the limiting case $\sigma_\tau = 0$. Results are shown in left two columns of Table 5.1. The increase in the variance of the delay corresponds to a higher fraction of complex eigenvalues; in particular, at the limiting cases, going from constant delay to the biggest variance considered, the percentage of complex eigenvalues goes from 52 to 61% and from 66 to 75%, in the two conditions.

We also study the effect of changing the mean of the delay, $\mu_\tau$, by considering three cases, in that we assume that all cells have a fixed delay $\tau$ of 0, 60 or 120 minutes; notice that $\tau = 0$ corresponds to replacing the delay term with $Z_t^N$. The impact of this change is dramatic: if we remove the delay, all eigenvalues have real solutions. Conversely, by increasing the delay to 1 hour, 98-99% of the eigenvalues have complex solutions only, and we reach 100% with a 2 hour delay.

We repeat the stability analysis, for each cell, using the parameters' posterior modes. Results are reported in the two right columns of Table 5.1. As expected, they are very similar to the ones discussed above. Figure 5.1 shows the function $a + b\, E_\tau(e^{-\lambda\tau})$ for basal and stimulated conditions. The points where the black

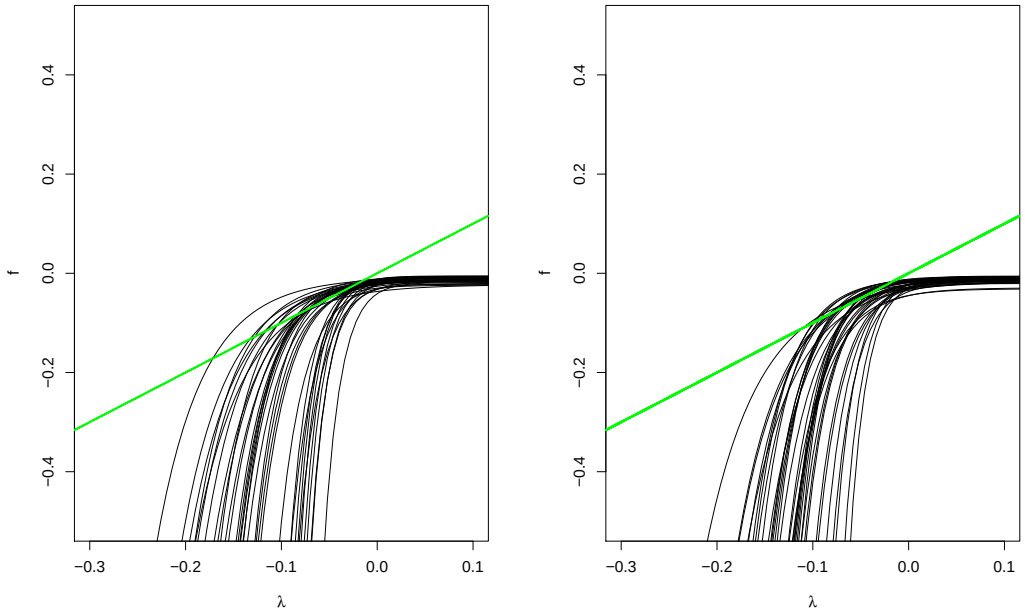


**Figure 5.1:** Plots for $a + b\, E_\tau(e^{-\lambda\tau})$ (black curves) and $\lambda$ (green 45° line), for the posterior modes of parameters under the basal (left image) and stimulated (right image) conditions. Each line corresponds to a cell.

curve crosses the 45° green line, indicating $\lambda$, represent the real solutions of the eigenvalues, if present. Clearly, not all curves intersect the green line: these cases correspond to complex eigenvalues, which have no solution on the real domain. In Figure 5.2, we also show how these plots vary when assuming a 1 hour delay for all cells, $\tau = 60$, showing that all curves are now below the green line, indicating no eigenvalue has a real solution.

The results presented above indicate that the delay term plays a key role in the stability of the Nrf2 system and suggest oscillations would be absent without it, even in a stochastic context. In particular, the mean value of the delay appears to have a stronger impact than the variability of its distribution. Recall, from Section 2.4, that the delay term represents a proxy for the nuclear phosphorylated Nrf2. Therefore, this result highlights the importance of the Nrf2 nuclear phosphorylation, by means of $a$Fyn, in the noise-induced oscillations.

Furthermore, we investigate the effect of the MM term by replacing it with a constant factor, $\theta_{MM} = \dfrac{k_a\, x_N^*}{K_a + x_N^*}$, where $Z_t^N$ is substituted by its stationary value;

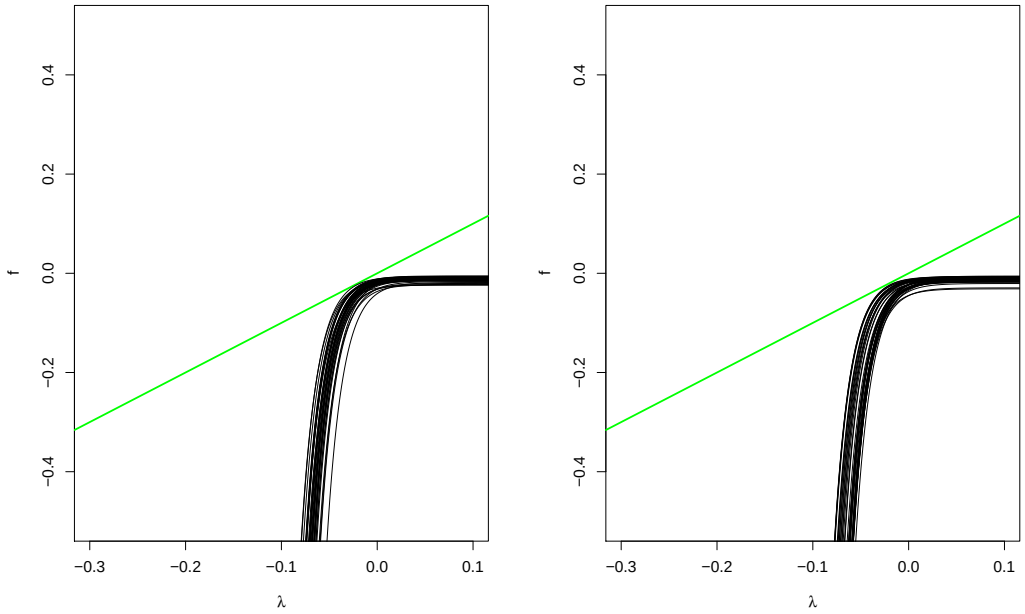


**Figure 5.2:** Plot for $a + b\,E_\tau(e^{-\lambda\tau})$ (black curves) and $\lambda$ (green 45° line), assuming a constant delay $\tau$ of 60 minutes, for the posterior modes of parameters under the basal (left image) and stimulated (right image) conditions. Each line corresponds to a cell.

the DDE system in (5.1)-(5.2) becomes

$$\dot{Z}_t^N = k_d\, Z_t^C - E_\tau(Z_{t-\tau}^N)\,\theta_{MM} - \delta Z_t^N \qquad [\text{ Nuc Nrf2 }] \tag{5.12}$$

$$\dot{Z}_t^C = E_\tau(Z_{t-\tau}^N)\,\theta_{MM} - k_d\, Z_t^C + \gamma - \delta Z_t^C. \qquad [\text{ Cyt Nrf2 }] \tag{5.13}$$

In this case, the system is already linear and does not need to be linearised. We follow a similar procedure as above, focusing on $Z_t^N$. We assume $Z_t^{cell}$ constant and express $Z_t^N$ as $Z_t^N = x_N^* + w_t$; hence we can rewrite (5.12) as

$$\dot{Z}_t^N = k_d\,(x_C^* - w_t) - (x_N^* + E_\tau(w_{t-tau}))\,\theta_{MM} - \delta(x_N^* + w_t) \tag{5.14}$$

$$\simeq b^* + w_t\,(-k_d - \delta) - E_\tau(w_{t-\tau})\,\theta_{MM} \tag{5.15}$$

$$\simeq w_t\,(-k_d - \delta) - E_\tau(w_{t-\tau})\,\theta_{MM}, \tag{5.16}$$

where $b^* = k_d\,x_C^* - x_N^*\,\theta_{MM} - \delta x_N^*$ is the DDE in 5.12 computed at equilibrium, which equals 0.

Moreover, by using, as before, $\dot{Z}_t^N = \dot{w}_t \simeq \lambda\, w_t$ and $w_t = w_0\, e^{\lambda t}$, we obtain

an equation for $\lambda$ as follows,

$$\lambda\, w_0\, e^{\lambda t} \simeq w_0\, e^{\lambda t}\,(-k_d - \delta) - w_0\, E_\tau(e^{\lambda(t-\tau)})\,\theta_{MM} \tag{5.17}$$

$$\lambda \simeq (-k_d - \delta) - E_\tau(e^{-\lambda\tau})\,\theta_{MM} \tag{5.18}$$

$$\lambda \simeq \tilde{a} + \tilde{b}\, E_\tau(e^{-\lambda\tau}), \tag{5.19}$$

which has the same form as before, where now coefficients are $\tilde{a} = -k_d - \delta$ and $\tilde{b} = -\theta_{MM}$.

The percentage of complex eigenvalues in this case, reported in the line corresponding to "Linear export" in Table 5.1, is only marginally lower than in the original non-linear system. However the small difference observed could be explained by the fact that, even in the non-linear system, we only consider a Taylor linear approximation of the MM and not the original non-linear term.

As before, we also repeat the analysis on the parameters' posterior modes of each cell and obtain similar results.

The results presented in this Section suggest a key component driving Nrf2 noise-induced oscillations is the delay term and in particular its mean value, while the variance of its distribution and the MM term play secondary roles. Finally, we find that the induction of cells by a stimulant is also a main factor in this process. This finding is coherent with the results shown in Section 4.6, where we show that induced cells correspond to higher import and export rates, and hence to faster and more dynamic movements between nucleus and cytoplasm.

## 5.4 DDE simulation with noise

In the previous section we found that, in a consistent fraction of cases, the Nrf2 system behaves like a noise-induced oscillator. In order to further investigate this phenomenon, we simulate from the DDE in (5.1)-(5.2), via the Euler method (Wilkinson, 2012), adding a random noise component at each time step. For simplicity, the analysis presented in this Section is only carried out on the posterior modes of parameters, therefore each cell contributes to 1 simulation only.

At every step of the Euler algorithm, we add correlated random noise in nuclear and cytoplasmic components, where the correlation between the two sources of noise is introduced to increase model realism. For each cell, we compute, via the DA covariance in (3.5), the correlation between nuclear and cytoplasmic increments, for the posterior mode parameter values and with $Z_t^N$ and $Z_t^C$ set equal to their equilibrium values. A histogram for the correlation values across the 71 available

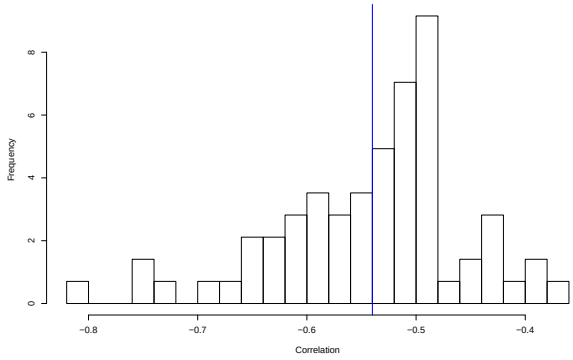


**Figure 5.3:** Correlation between nuclear and cytoplasmic variations, computed from the DA covariance matrix on each one of the available 71 cells, with parameter values set to their posterior modes and with $Z_t^N$ and $Z_t^C$ equal to their equilibrium values. The blue vertical line denotes the average correlation, $-0.54$.

cells is shown in Figure 5.3. The average correlation, over all 71 cells, between nuclear and cytoplasmic variations is $-0.54$.

At every Euler step, of time interval equal to 1 minute, we add a white noise normal component $\epsilon^* \sim \mathcal{N}(0, \sqrt{\Delta}\,\Sigma_\epsilon^*)$, with $\Sigma_\epsilon^* = \begin{pmatrix} 1 & -0.54 \\ -0.54 & 1 \end{pmatrix}$, with $\Delta = 1$ minute being the Euler step. The variance is chosen to be 1 because this value allows us to reproduce movements of a realistic size. Increasing of decreasing the variance produces analogous results to the ones described below.

We simulate each process for $10^4$ minutes: all simulations quickly converge to their stationary values and fluctuate around them. Figure 5.4 shows an example of a simulated process for a cell: the top images refer to the entire $10^4$ window, whilst the bottom panels represent a zoomed version over 400 minutes, which is a similar window to the time length of the experimental data. Indeed, these plots show very similar patterns to the ones exhibited by the available data. Simulations for all cells behave qualitatively similar to the one displayed in Figure 5.4. When decreasing the variance matrix $\Sigma_\epsilon^*$, by factors from 10 to $10^4$, we obtain analogous plots, with smaller deviations around equilibrium.

Images of nuclear versus cytoplasmic levels, in Figures B.44-B.49 in the Appendix, show how each simulation spirals around its stationary solution, represented by the green dot at the centre of the vortex. Figure 5.5 shows the same image for the two cells we have taken as example so far: cell 4 under the basal conditions and cell 3 under the stimulated one. On the left image we see how the simulation converges towards its equilibrium, denoted by the green dot, whilst the right plots show a

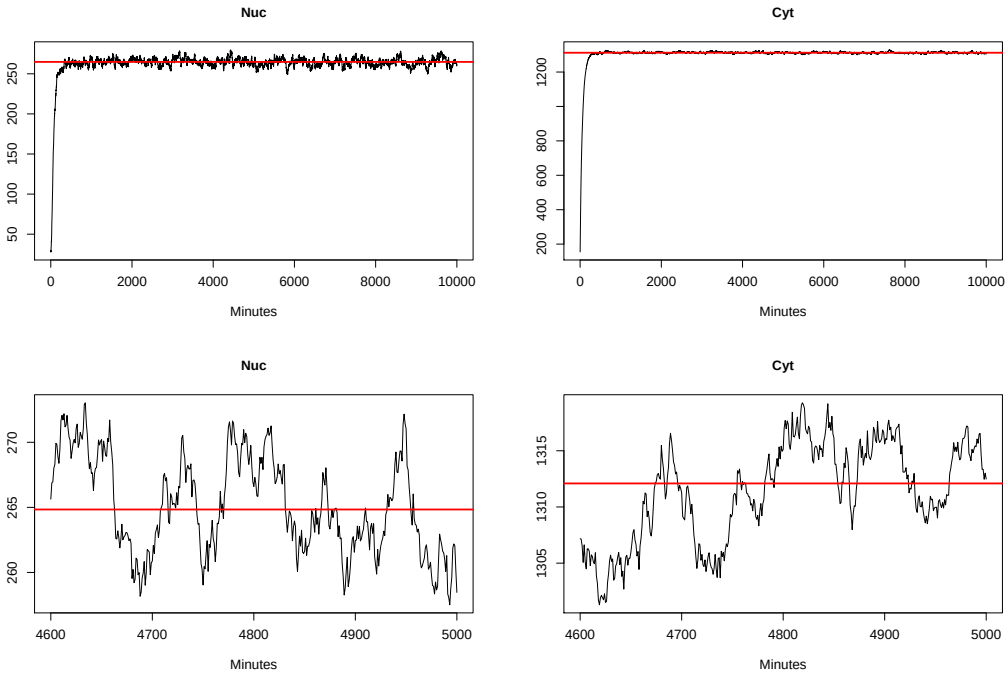


**Figure 5.4:** Example of an Euler simulation, with random noise added, for nuclear (left images) and cytoplasmic (right images) Nrf2 reporter population of molecules; the red horizontal line indicates the stationary solution of the DDE. The top row shows the entire simulation, over 10,000 minutes, whilst the bottom row displays a 400 minute zoom in the central part of the simulation. Parameters used are the posterior mode of parameters from cell number 10 under the basal condition; in particular they correspond to: $k_d = 6.15\,10^{-3}$, $k_a = 1.38\,10^{-2}$, $K_a = 1.52\,10^{-1}$, $\mu_\tau = 18.24$, $\gamma = 26.36$ and $\delta = 1.67\,10^{-2}$.

zoomed version of the same simulation, without convergence, around the stationary point only.

Furthermore, we compute, from each simulated process, the autocorrelation (ACF) plot to get a qualitative measure of the oscillatory behaviour: as with the experimental data, ACF plots vary greatly between cells and, in some cases, exhibit patterns associated with strong oscillations, like the ones shown in Figure 5.6. Autocorrelation plots for all simulated processes can be found in the Appendix, in Figures B.50-B.61. Furthermore, Figures B.62-B.73 show all ACFs for the experimental data, which exhibit similar patterns to the ones shown by the simulated data.

Table A.4 reports the fraction of complex eigenvalues amongst the 100 simulated processes in each cell. Comparing these numbers with the ACFs from, both experimental and simulated data, we notice that ACFs corresponding to cells with a very high proportion of complex eigenvalues are more strongly oscillatory than ACFs

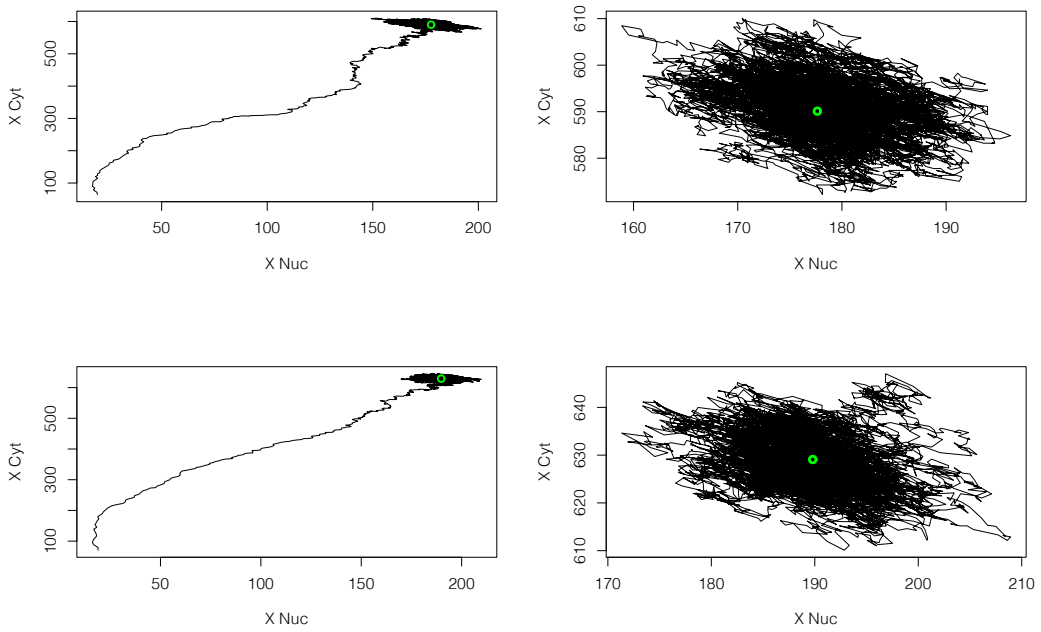


**Figure 5.5:** Graphical representation, with nuclear and cytoplasmic processes on the two axes, of Euler simulations perturbed with noise; each simulation moves around the respective DDE solution, represented by the green dot. The top two images refer to the process simulated from the posterior mode of parameters for cell 4 under the basal condition; the bottom two images show to the process simulated from the posterior mode of parameters for cell 3 under the stimulated condition. On the left plots the entire process is represented; the right images show a zoom version of the same plots around equilibrium, removing convergence to it.

for cells with mostly real eigenvalues. This confirms the close connections between oscillatory patterns and complex eigenvalues and it shows how, by introducing a source of white noise, a significant proportion of simulations display oscillatory patterns around the stationary solutions.

We compare the ACFs in Figure 5.6 with those obtained from our experimental data. Figure 5.7 shows the ACFs for the same two cells as in Figure 5.6, which also exhibit strong periodicity that is normally associated with oscillatory patterns. There is a high degree of similarity between the ACFs for real data and the ones for the DDE noise perturbed simulation presented here; the analogy is not only present on the two examples in Figures 5.6 and 5.7 yet also on the remaining cells. For instance the regular periodicities shown here are only observed in a minority of cases in both real and simulated data.

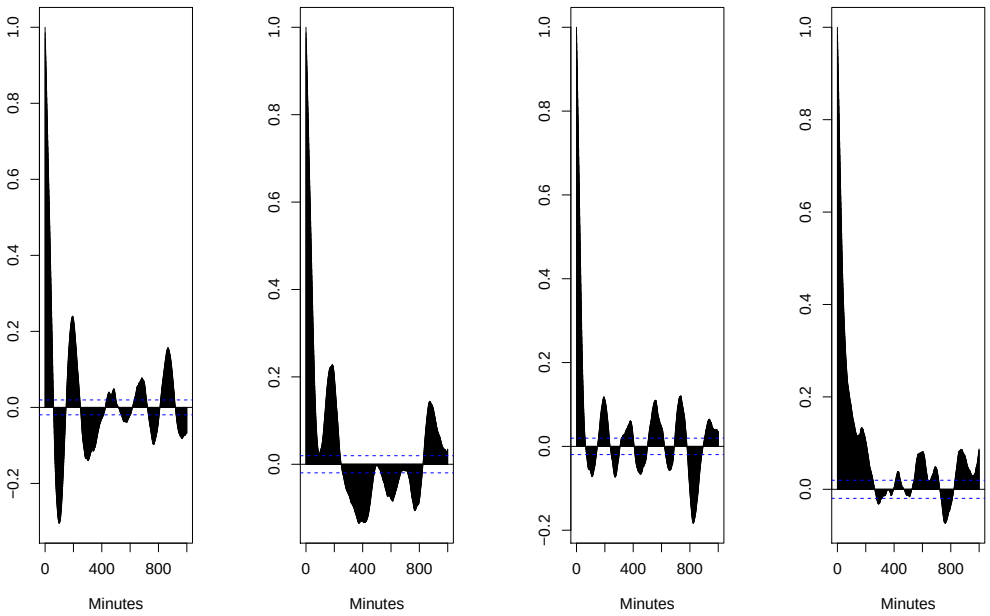


**Figure 5.6:** Autocorrelation plot, of lag 1-1,000 minutes, for two simulated processes. The left two images refer to nucleus and cytoplasm, respectively, simulated from posterior mode of parameters for cell 4 under the basal condition; the right two images, to nuclear and cytoplasmic processes simulated from the posterior mode of parameters for cell 3 under the stimulated condition.

The ACFs referring to the simulated data show more fluctuations due to longer lag the autocorrelation is computed for, which is 1,000 minutes; this is possible because of the long time the process is simulated for. The real data ACF, instead, is limited to the time length the cell is observed for.

Images in Figures 5.6 and 5.7 correspond to cell number 4 under the basal condition and to cell number 3 under the stimulated condition; the original processes for these cells, visible in Figures B.1 and B.6 in the Appendix, are strongly oscillatory.

Hence, the ACF provides reassuring evidence that the parameters we infer in oscillatory cells are then able to reproduce the oscillatory dynamics observed. Moreover, considering the eigenvalue study in the previous Section, the basal and stimulated cells considered here, i.e. basal cell number 4 and stimulated cell number 3, have complex eigenvalues in 100 and 97% of the case, respectively, as visible in Table A.4. This confirms how, in oscillatory cells, our methodology mostly infers parameter values which correspond to complex eigenvalues; those parameters, in stochastically perturbed simulations, are also able to reproduce oscillations.

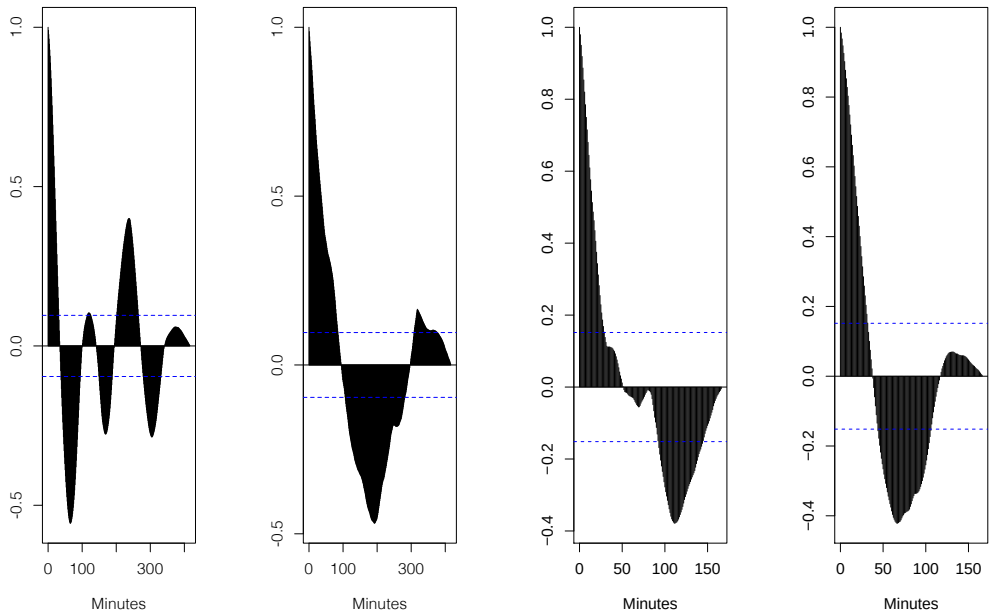


**Figure 5.7:** Autocorrelation plot for two cells from the available experimental data. The left two images refer to nucleus and cytoplasm, respectively, for the real process of cell 4 under the basal condition; the right two images, to nuclear and cytoplasmic real processes for cell 3 under the stimulated condition.

We also point out that very similar results, in terms of both time series plots and ACFs, are obtained when simulating from the stochastic DA model, in particular from its EM approximation, following the steps described in Section 3.7. This is a further indication that stochastically perturbing the DDE of our system, we achieve processes comparable to the ones obtained via its stochastic counterpart.

## 5.5 Summary

In this Chapter we performed a stability analysis of the Nrf2 system around its equilibrium. Results show that, for a large number of cells, the Nrf2 system behaves like a noise-induced oscillator which, under the deterministic setting, converges to a stable limit point solution and, when perturbed by white noise, displays oscillations. The stimulation increases the proportion of cells with this characteristic: the fraction of posterior parameters leading to complex eigenvalues increases from 53.0% to 66.5% in stimulated cells, as compared to cells under the basal condition. Furthermore, complex eigenvalues are more frequent in cells characterized by oscillatory patterns,

which provides reassuring evidence that, in oscillatory cells, our methodology infers parameter values which are able to reproduce oscillations.

We also find that the delay term represents a key component of the oscillatory mechanism; in particular, increasing the mean delay to 1 hour results in almost all eigenvalues having complex solutions. The delay term models the amount of phosphorylated nuclear Nrf2 and controls the delayed nuclear export signal. This result suggests that such a phosphorylation, triggered by $a$Fyn, plays a central role in the Nrf2 oscillatory mechanism and in its intracellular movements.

Finally, we have shown that, by perturbing the DDE with a source of random noise, we obtain qualitatively very similar oscillations to those displayed by our experimental data, as seen by inspecting the time series plots and empirical ACFs. Analogous results are obtained when simulating from its stochastic counterpart, via the DA and EMA.

# Part II

# Inference on single cell gene expression from a two-state switch gene model

# Chapter 6

# Biology

> A ship in harbor is safe, but that is not what ships are built for.
>
> *John Augustus Shedd*

This Part of the thesis focuses on a different biological problem than the one considered in Part I. However, both analyses are based on dynamical stochastic models for molecular populations in single cells and also share some inferential tools, such as the Bayesian hierarchical framework, the data augmentation procedure and the Metropolis-within-Gibbs sampling structure.

This Chapter presents an introduction to the biological background, the motivation of the study and the available data, which will be used for inference. It also describes the modelling approach, namely a two-state transcription switch model which is assumed for the gene of interest, and compares it with two, less realistic, alternative models.

In this Part of the thesis there is no reference to any notation used in Part I, hence all notation needed will be redefined.

## 6.1 Motivation

The analysis focuses on investigating expression of messenger RNA (mRNA), in single cells, from a large population of cells. Transcription, i.e. the process in which mRNA transcripts are synthesized from genes, clearly is an inherently stochastic phenomenon (Singh *et al.*, 2013): mRNA levels vary greatly between cells, even for genetically identical cells under the same experimental and environmental conditions (Hebenstreit, 2013; Kim and Marioni, 2013; Raj *et al.*, 2006). Models of the way gene

expression is regulated involve and depend upon a series of interactions of biological elements and events, whose timing is typically random (Shahrezaei and Swain, 2008; Kaern *et al.*, 2005; Kim and Marioni, 2013). For instance, transcription is regulated by factors binding at upstream promoter elements; these binding events are the result of random encounters of molecules and hence contribute to the inherent stochasticity of transcription (Kaern *et al.*, 2005). Investigating biological noise is of particular interest as it could lead to an improved understanding of these cellular mechanisms. In order to do so, we develop a stochastic dynamic model for the mRNA molecular population in single cells and derive its stationary solution, which we fit to multiple single cell data obtained via fluorescence in situ hybridization (FISH) flow cytometry (FC) experiments.

We propose a two-state switch model where the gene alternates between a more active ON state and a less active OFF state. Despite the use of the word "off", we note that mRNA may also be transcribed in the OFF state, although transcription is much higher while the gene is ON. This structure is particularly appealing because, in spite of its relative simplicity, it allows the modelling of transcriptional bursts, corresponding to relatively short periods of time where high quantities of mRNA are transcribed. This phenomenon is typical of many genes and species (Dar *et al.*, 2012; Golding *et al.*, 2005; Raj *et al.*, 2006; So *et al.*, 2011; Suter *et al.*, 2011; Zopf *et al.*, 2013; Rajala *et al.*, 2010), although its underlying mechanism is largely unclear.

In order to improve the current knowledge about this transcription mechanism, we analyse gene expression data in single cells, from a gene whose transcription is believed to happen in bursts. Data are collected in cell culture, for a version of the human immunodeficiency virus type 1 (HIV-1) *env* gene, under several experimental conditions, namely at different levels of induction by the stimulant *tetracycline*, which boosts gene expression levels. We are particularly interested in understanding the effect that stimulation has on the system.

In the next Sections we will describe in more detail the two-state switch model we employ as well as the available experimental data.

## 6.2 Data

The cell lines we use have been given to the Hebenstreit lab at Warwick by Mapendano *et al.* (2010) and are a modified version of HEK293 cells containing a version of the HIV-1 *env* gene under the control of a *tetracycline* inducible promoter.

The mRNA levels are observed, in each cell separately, via fluorescence in situ hybridization (FISH) flow cytometry (FC); the native mRNA is tagged with fluores-

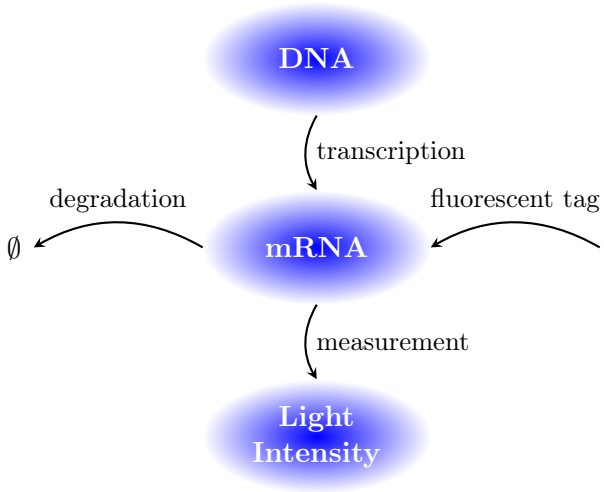


**Figure 6.1:** Measurement process for the mRNA from the HIV-1 *env* gene.

cent labelled oligos, which are short nucleotide sequences designed to bind specifically to the mRNA of interest. A laser is then used to induce these tagged mRNAs to emit light; the measurement procedure is illustrated in Figure 6.1. Therefore measurements actually refer to the original mRNA, which comes from the unmodified DNA, not to a reporter as with the Nrf2 analysis in Part I. The software used for the analysis, the BD FACSDiva$^{\text{TM}}$ software of the BD LSRFortessa$^{\text{TM}}$ cell analyzer, measures the overall light intensity in each cell and converts it into a number. In every experiment, we record about 10,000 observations, each one from a different cell. Every cell corresponds to one measurement, which is assumed to come from the stationary distribution of HIV-1 *env* mRNA expression.

Our data consist of several sets, each observed under a distinct experimental condition, i.e. at different levels of induction by *tetracycline*: 0, 5 and 10 ng/ml (nanograms per millilitre), where 0 clearly indicates no stimulation.

For each experimental condition, we repeat the experiment four times; hence, we have four replicate datasets which we treat hierarchically. This allows us to explicitly model the experimental variability between replicates and use hyperparameters to carry out comparisons between different experimental conditions.

## 6.3 Switch Model

One of the most basic models for gene expression assumes that, in each cell, transcription and degradation of mRNA molecules occur as a birth and death process with exponential waiting times, with constant rates that we call $\alpha$ and $\beta$, respec-

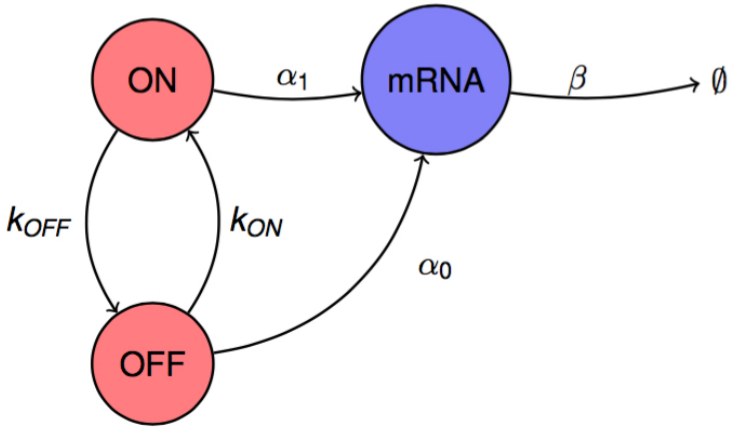


**Figure 6.2:** System for the two-state switch model we propose. The red circles indicate the state of the gene, ON or OFF, whilst the blue circle represents the mRNA population.

tively. If we define $X_t$ as the population of mRNA at time $t$ in a cell, we find that the stationary distribution of mRNA is Poisson distributed(Paulsson, 2005; Singh *et al.*, 2013; Raj *et al.*, 2006), in particular $X_t \sim Pois\left(\frac{\alpha}{\beta}\right)$ (Munsky *et al.*, 2012). Such a simplistic model typically under-estimates the variability of the real biological mechanism and, hence fails to capture the main features of the data. In fact, the distribution of gene expression is normally over-dispersed with respect to the Poisson, with variance significantly larger than the mean. In particular, this model usually fails to capture the broad mRNA distributions from regulated genes (Munsky *et al.*, 2012).

A more realistic assumption consists of using a two-state switch model which assumes the gene stochastically alternates between ON and OFF states, with exponentially distributed waiting times, with constant rates that we call $k_{ON}$, for the change from OFF to ON, and $k_{OFF}$, for the change from ON to OFF. The gene only transcribes mRNA in the ON state, while in the OFF transcription is assumed to be null (Suter *et al.*, 2011; Wills *et al.*, 2013; Hebenstreit, 2013; Kim and Marioni, 2013; Munsky *et al.*, 2012; Peccoud and Ycart, 1995); degradation still happens, as above, at constant rate $\beta$. The resulting stationary distribution is the heavier tailed Poisson-beta distribution, which is an over-dispersed Poisson where one of the components of the parameter is a beta distributed random variable (Johnson *et al.*, 2005; Kim and Marioni, 2013). This assumption highly improves realism, with respect to the simple one-state model we introduced above, and allows us to model transcriptional bursts.

Nevertheless, assuming no transcription in the OFF state could be an unre-

alistic assumption. Hence, we propose a two-state switch model similar to the one above, yet where transcription also occurs at a less active rate in the OFF state. The gene transcription still is assumed to switch between ON and OFF states with exponentially distributed waiting times, at rates $k_{ON}$, when moving from OFF to ON, and $k_{OFF}$, when going from ON to OFF. The gene transcribes mRNA at two distinct rates, corresponding to either rate, which we call $\alpha_1$ and $\alpha_0$, respectively, with $\alpha_1 > \alpha_0$; therefore, while in the OFF state, the gene is less active yet not completely dormant. Degradation still happens at the constant per molecule rate $\beta$, regardless of the gene condition. Figure 6.2 graphically illustrates the process.

As we will show in Section 7.1, the mRNA population at equilibrium is still distributed as a Poisson-beta, although clearly with different parameters than in the simpler model, where $\alpha_0 = 0$. The resulting distribution is a key element in inference, as it allows us to easily sample the mRNA population at stationarity and to compute its density. Notice that the previous two-state switch model, with no transcription in the OFF state, is a particular case of ours, with $\alpha_0 = 0$. Also the simpler one-state model can be obtained from ours, assuming that the gene is constantly ON or OFF, i.e. $k_{OFF} = 0$ or $k_{ON} = 0$.

Figure 6.3 shows an example of densities for the mRNA cellular population under the three models we introduced; all cases share the same mean, 44.5, and the two-state switch models have the same switch coefficients, $k_{ON}$ and $k_{OFF}$. Parameter values are chosen, after analysing the experimental data, in order to roughly match their average value and typical shape.

We notice that the one-state model, in red, leads to a peaked Poisson distribution. The two-state switch model with $\alpha_0 = 0$, in blue, is highly asymmetric, with very long right tail, and the mode is quite distant from the mean. The model we propose, in black, is less asymmetric but can account for overdispersion, in contrast to the Poisson.

## 6.4 Reaction network

We introduce a reaction network to our two-state switch model: we call $S(t)$, or simply $S_t$, the state of the gene at time $t$, either 0 (OFF) or 1 (ON), and $X(t)$, or $X_t$, the numbers of mRNA molecules at time $t$ in a single cell.

The model we propose can be represented by 5 reactions, $R_1, ..., R_5$, for the bivariate populations $(X_t, S_t)^T$. The $i$-th reaction is associated to a stoichiometry vector, which we call $v_i$, that indicates the changes generated by $R_i$ in the populations of species, for $i = 1, ..., 5$. Furthermore, $R_i$ is also associated to a hazard,

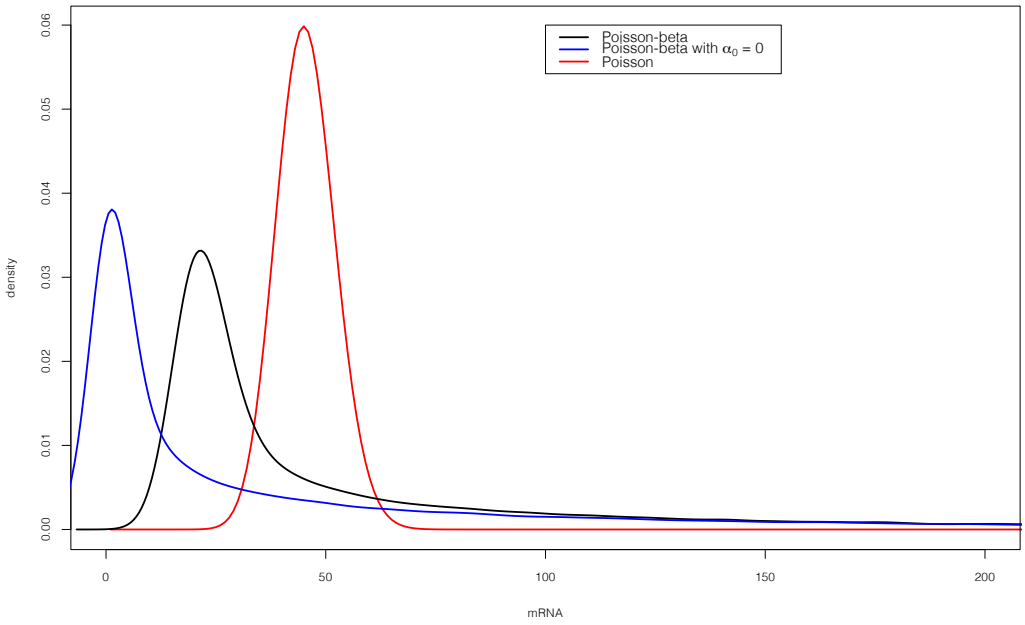


**Figure 6.3:** Example of the densities of the number of mRNA molecules in single cells for the one-state model (in red), with mean parameter 44.5, the two-state switch model with transcription in the ON state only (in blue) , with parameters $\alpha_1 = 1,780$, $k_{ON} = 2.5$ and $k_{OFF} = 97.5$ and the two-state switch model we propose with transcription in both states (in black), with parameters $\alpha_1 = 1,000$, $\alpha_0 = 20$, $k_{ON} = 2.5$ and $k_{OFF} = 97.5$. All cases have the same mean, 44.5.

defined as $w_i(X_t, S_t, \theta)$, that depends on the population values, at time $t$, and on a vector containing the parameters of interest, $\theta = (k_{ON}, k_{OFF}, \alpha_1, \alpha_0, \beta)$. Given an infinitesimal time $dt$, $w_i(X_t, S_t, \theta)\, dt$ represents the probability that reaction $R_i$ occurs in the time interval $(t, t + dt]$, $i = 1, ..., 5$ (Wilkinson, 2012).

In the following we introduce the 5 reactions and corresponding stoichiometry vectors and hazards.

- $R_1$ represents the gene switching from the OFF to the ON state, i.e. $S_t$ changes from 0 to 1, hence $v_1 = (0, 1)$. The hazard is $w_1(X_t, S_t, \theta) = k_{ON}\,(1 - S_t)$, where $(1 - S_t)$ plays the role of a binary indicator, which is 0 if the gene is already ON and 1 if the gene is OFF; this allows us to model the fact that the gene can only be turned ON when it is in the OFF state.

- $R_2$ indicates the gene switching from the ON to the OFF state, in this case $S_t$ changes from 1 to 0, with $v_2 = (0, -1)$. The hazard is $w_2(X_t, S_t, \theta) = k_{OFF}\, S_t$,

where $S_t$ is 1 if the gene is ON and 0 if the gene is already OFF; again, the hazard is non-null only if the gene is ON.

- $R_3$ corresponds to the transcription of 1 molecule of mRNA, while in the ON state, with rate $w_3(X_t, S_t, \theta) = \alpha_1 S_t$. The stoichiometry vector is $v_3 = (1, 0)$.
- $R_4$ again represents the transcription of 1 molecule of mRNA, i.e. $v_4 = (1, 0)$, this time in the OFF state, with rate $w_4(X_t, S_t, \theta) = \alpha_0 (1 - S_t)$.
- $R_5$ indicates the degradation of 1 molecule of mRNA, regardless of the state of the model, with linear rate $w_5(X_t, S_t, \theta) = \beta X_t$; hence i.e. $v_5 = (-1, 0)$.

This reaction network is summarized in Table 6.1.

| Reaction | $R_1$ | $R_2$ | $R_3$ | $R_4$ | $R_5$ |
|---|---|---|---|---|---|
| Reaction description | Gene turns ON | Gene turns OFF | Transcription ON state | Transcription OFF state | Degradation |
| Change | $S_t \to S_t + 1$ | $S_t \to S_t - 1$ | $X_t \to X_t + 1$ | $X_t \to X_t + 1$ | $X_t \to X_t - 1$ |
| $v_i$ | (0, 1) | (0, -1) | (1, 0) | (1, 0) | (-1, 0) |
| $w_i(X_t, S_t, \theta)$ | $k_{ON}(1 - S_t)$ | $k_{OFF} S_t$ | $\alpha_1 S_t$ | $\alpha_0 (1 - S_t)$ | $\beta X_t$ |

**Table 6.1:** Reactions included in the two-state switch gene model reaction network and respective hazards.

## 6.5 Summary

In this Chapter we have introduced the biological background for the analysis of gene expression from a two-state switch model. We have presented the two-state switch model with transcription in both ON and OFF states, as well as two simpler and less realistic alternatives. We have also described the reaction network associated with this model.

In the next Chapter we will introduce the mathematical details of the system, we will prove that its stationary distribution can be seen as a Poisson-beta and we will describe the inferential procedure that will be applied to the experimental data.

# Chapter 7

# Statistical methods

> Beware that, when fighting monsters, you yourself do not become a monster...for when you gaze long into the abyss, the abyss also gazes into you.
>
> *Friedrich Nietzsche*

This Chapter presents original material proving that the two-state switch model, with transcription in both states, i.e. $\alpha_1, \alpha_0 \geqslant 0$, has a Poisson-beta stationary distribution for the mRNA population. We derive the measurement equation relating the observations to the unobserved population levels, which are treated as latent states that we sample via a data augmentation procedure. We describe how multiple replicates can be embedded in a Bayesian hierarchical framework and the MCMC algorithm used for sampling the posterior distributions of the parameters. Finally, we illustrate simulation studies checking the performance of our inferential procedure.

## 7.1 Poisson-Beta Derivation

In this Section we present the explicit density for $X_t$ when the cell is at equilibrium, developed by Singh *et al.* (2013), and we show that such mRNA stationary distribution corresponds to a Poisson-beta. The latter equivalence has been proved in unpublished notes by Lucy Ternant, former MOAC Master and first year PhD student at the University of Warwick under the supervision of Dr Daniel Hebenstreit and Prof. Bärbel Finkenstädt.

In contrast to the analysis in Part I of this thesis, which relies upon an approximate likelihood of the system, the density we will present here is the exact stationary density of the model in Table 6.1, under the assumption that the mRNA population in each cell has converged to an equilibrium, such that the mRNA distribution over the cells is identical to the stationary distribution. The available data are recorded after 16 hours from induction.

Singh *et al.* (2013) show that the reaction network in Table 6.1 has the following stationary distribution for the mRNA count at equilibrium,

$$P(X=x) = e^{-\tilde{\alpha}_0} \sum_{i=0}^{x} \left( \frac{\Gamma(\tilde{k}_{ON}+i)\,\Gamma(\tilde{k}_{ON}+\tilde{k}_{OFF})}{\Gamma(\tilde{k}_{ON})\,\Gamma(\tilde{k}_{ON}+\tilde{k}_{OFF}+i)} \frac{\tilde{\alpha}_0^{x-i}}{(x-1)!} \times \frac{(\tilde{\alpha}_1-\tilde{\alpha}_0)^i}{i!}\, {}_1F_1(\tilde{k}_{ON}+i;\tilde{k}_{ON}+\tilde{k}_{OFF}+1;\tilde{\alpha}_0-\tilde{\alpha}_1) \right), x \in \mathbb{N}, \tag{7.1}$$

where $P(.)$ indicates the probability operator, $\Gamma$ refers to the gamma function, ${}_1F_1$ is the confluent hypergeometric function of the 1st kind and all parameters are rescaled with respect to the degradation rate $\beta$: $\tilde{k}_{ON} = \frac{k_{ON}}{\beta}$, $\tilde{k}_{OFF} = \frac{k_{OFF}}{\beta}$, $\tilde{\alpha}_1 = \frac{\alpha_1}{\beta}$ and $\tilde{\alpha}_0 = \frac{\alpha_0}{\beta}$. Parameters $k_{ON}$, $k_{OFF}$, $\alpha_1$, $\alpha_0$ and $\beta$ have been defined in Sections 6.3 and 6.4. Therefore, as in the simpler two-state switch model with $\alpha_0$, the parameters are normalized with respect to the degradation rate (Kim and Marioni, 2013; Shahrezaei and Swain, 2008). As a consequence, the time unit of these four parameters also changes as they are now expressed in terms of per molecule degradation rate; e.g. to obtain the per minute transcription rate in the ON state, we should multiply parameter $\tilde{\alpha}_1$ by the per minute degradation rate. Parameter $\beta$ only appears in the stationary density as a linear combination with the other four parameters, hence it is not structurally identifiable. Our inference will therefore focus on the four standardized parameters $\tilde{k}_{ON}, \tilde{k}_{OFF}, \tilde{\alpha}_1$ and $\tilde{\alpha}_0$.

*Theorem:* The density in (7.1) can be associated with the following latent variable structure,

$$X = A + B \tag{7.2}$$

$$A \sim Pois((\tilde{\alpha}_1 - \tilde{\alpha}_0)\,P) \tag{7.3}$$

$$P \sim Beta(\tilde{k}_{ON}, \tilde{k}_{OFF}) \tag{7.4}$$

$$B \sim Pois(\tilde{\alpha}_0), \tag{7.5}$$

where the probability functions for $A$ (Johnson *et al.*, 2005) and $B$ are

$$\begin{aligned} P(A=y) &= \frac{\Gamma(\tilde{k}_{ON}+y)\,\Gamma(\tilde{k}_{ON}+\tilde{k}_{OFF})}{\Gamma(\tilde{k}_{ON})\,\Gamma(\tilde{k}_{ON}+\tilde{k}_{OFF}+y)} \times \\ &\quad \frac{(\tilde{\alpha}_1-\tilde{\alpha}_0)^y}{y!}\,{}_1F_1(\tilde{k}_{ON}+y;\tilde{k}_{ON}+\tilde{k}_{OFF}+1;\tilde{\alpha}_0-\tilde{\alpha}_1),\ y\in\mathbb{N}, \end{aligned} \tag{7.6}$$

$$P(B=z) = e^{-\tilde{\alpha}_0}\,\frac{\tilde{\alpha}_0^z}{z!},\ z\in\mathbb{N}. \tag{7.7}$$

*Proof:* Since $X$ is defined as the summation of $A$ and $B$, $X=x$ is obtained when $(A=a,B=b)$ with $(a,b)\in\{(0,x),(1,x-1),...,(x-1,1),(x,0)\}$. Furthermore, given $A$ and $B$ are independent $P(A=a,B=b)=P(A=a)\,P(B=b)$. Hence, the density for $X$ can be obtained, via the discrete convolution formula, as

$$P(X=x) = \sum_{i=0}^{x} P(A=i)\,P(B=x-i), x\in\mathbb{N} \tag{7.8}$$

$$\begin{aligned} &= e^{-\tilde{\alpha}_0}\sum_{i=0}^{x}\frac{\Gamma(\tilde{k}_{ON}+i)\,\Gamma(\tilde{k}_{ON}+\tilde{k}_{OFF})}{\Gamma(\tilde{k}_{ON})\,\Gamma(\tilde{k}_{ON}+\tilde{k}_{OFF}+i)}\,\frac{\tilde{\alpha}_0^{x-i}}{(x-1)!}\times \\ &\quad \frac{(\tilde{\alpha}_1-\tilde{\alpha}_0)^i}{i!}\,{}_1F_1(\tilde{k}_{ON}+i;\tilde{k}_{ON}+\tilde{k}_{OFF}+1;\tilde{\alpha}_0-\tilde{\alpha}_1), \end{aligned} \tag{7.9}$$

which corresponds to the formula in (7.1). Going from (7.8) to (7.9) we simply replace the probabilities of $A$ and $B$ with their formulae in (7.6) and (7.7), respectively. This completes the proof of the Theorem.

Therefore we have shown that $X$ can be written as the summation of $A$ and $B$, as in (7.2)-(7.5). Furthermore, since the summation of two independent Poisson random variables (r.v.s) still is a Poisson r.v., we can rewrite the distribution of $X$ as

$$X \sim Pois((\tilde{\alpha}_1-\tilde{\alpha}_0)\,P+\tilde{\alpha}_0), \text{ with } P\sim Beta(\tilde{k}_{ON},\tilde{k}_{OFF}) \tag{7.10}$$

or, equivalently,

$$X \sim Pois(\tilde{\alpha}_1\,P+\tilde{\alpha}_0\,(1-P)), \text{ with } P\sim Beta(\tilde{k}_{ON},\tilde{k}_{OFF}). \tag{7.11}$$

From this structure we can define the mean and variance of $P$, as

$$E(P) = \frac{\tilde{k}_{ON}}{\tilde{k}_{ON}+\tilde{k}_{OFF}} \tag{7.12}$$

$$Var(P) = \frac{\tilde{k}_{ON}\,\tilde{k}_{OFF}}{(\tilde{k}_{ON}+\tilde{k}_{OFF})^2\,(\tilde{k}_{ON}+\tilde{k}_{OFF}+1)} \tag{7.13}$$

As in the basic model with $\alpha_0 = 0$, $P$ can be interpreted as the probability that the gene is in the ON state (Johnson *et al.*, 2005), where $E(P)$ represents the average time the gene spends in the ON state.

We also describe the mean and variance of $X$. Considering equation (7.2) and owing to the independence of $A$ and $B$, we can write the mean and variance of $X$ as $E(X) = E(A) + E(B)$ and $Var(X) = Var(A) + Var(B)$. Clearly $E(B) = Var(B) = \tilde{\alpha}_0$ follows from the Poisson distribution of $B$. While, from Johnson *et al.* (2005), we know that:

$$E(A) = (\tilde{\alpha}_1 - \tilde{\alpha}_0) \frac{\tilde{k}_{ON}}{\tilde{k}_{ON} + \tilde{k}_{OFF}}$$

and

$$Var(A) = E(A) + \frac{\tilde{k}_{ON}\, \tilde{k}_{OFF}\, (\tilde{\alpha}_1 - \tilde{\alpha}_0)^2}{(\tilde{k}_{ON} + \tilde{k}_{OFF})^2\, (\tilde{k}_{ON} + \tilde{k}_{OFF} + 1)}.$$

From the equations above, we can easily obtain the following formulations for the mean and variance of X:

$$E(X) = \tilde{\alpha}_0 + (\tilde{\alpha}_1 - \tilde{\alpha}_0) \frac{\tilde{k}_{ON}}{\tilde{k}_{ON} + \tilde{k}_{OFF}} \tag{7.14}$$

$$Var(X) = E(X) + \frac{\tilde{k}_{ON}\, \tilde{k}_{OFF}\, (\tilde{\alpha}_1 - \tilde{\alpha}_0)^2}{(\tilde{k}_{ON} + \tilde{k}_{OFF})^2\, (\tilde{k}_{ON} + \tilde{k}_{OFF} + 1)} \tag{7.15}$$

The Poisson-beta structure for $X$ in (7.10), introduced in the Theorem above, highly facilitates inference as it easily allows one to sample the mRNA population at equilibrium. Furthermore ${}_1F_1$, which appears in (7.1), needs to be numerically estimated; this process is troublesome as there is no numerical method which is accurate, quick and reliable at the same time in the entire parameter space (Kim and Marioni, 2013; Muller, 2001). The methodology we propose circumvents the need to compute ${}_1F_1$, as it takes advantage of the latent variable structure in (7.10) to sample the mRNA levels without using their probability density $P(X)$.

An alternative sampling scheme might be to use the stochastic Gillespie algorithm (Gillespie, 1977) and simulate directly from the reaction network in Table 6.1, where the time to the next $k$-th reaction is exponentially distributed with rate equal to the $k$-th hazard, $k = 1, ..., 5$. Since the degradation rate is unidentifiable, one can assume $\beta = 1$ and, hence, use $\alpha_0 = \tilde{\alpha}_0$, $\alpha_1 = \tilde{\alpha}_1$, $\tilde{k}_{OFF} = k_{OFF}$ and $\tilde{k}_{ON} = k_{ON}$. Figure 7.1 displays an example of a Gillespie simulation for the mRNA population; particularly interesting is the right image which shows how the model simulates transcriptional bursts when the gene turns ON, corresponding to sudden

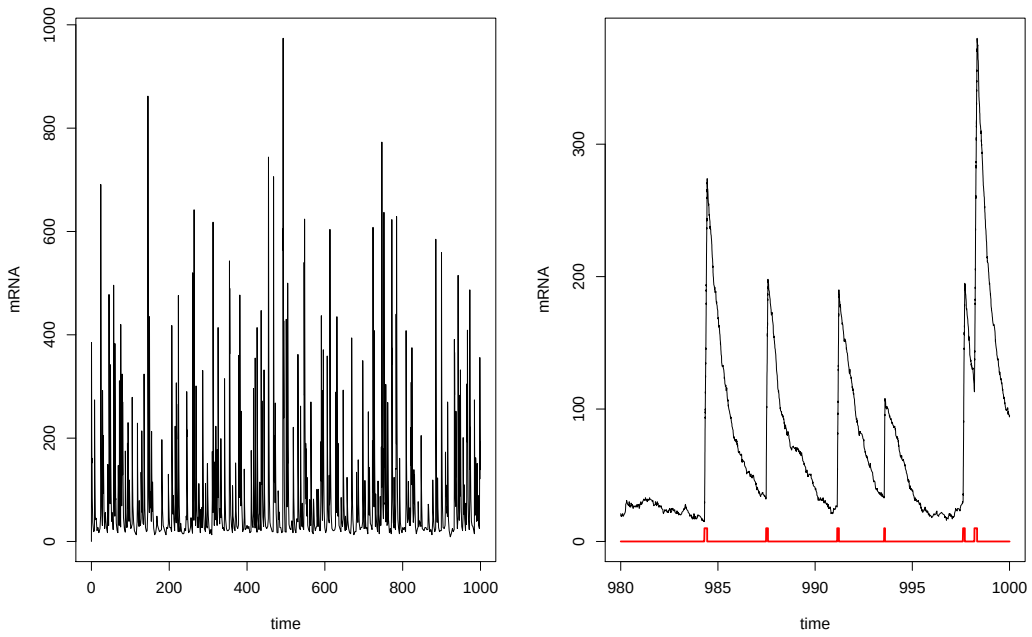


**Figure 7.1:** Example of a Gillespie simulation (Gillespie, 1977) of the number of mRNA molecules in single cells for the two-state switch model; entire simulation on the left and zoom on 20 minutes on the right. The bottom red line indicates the state of the gene, which is mostly OFF and turns ON when the line goes up vertically, corresponding to the moments when mRNA is transcribed at rate $\alpha_1$. Parameters used for the simulation are: $\alpha_0 = 20$, $\alpha_1 = 1,000$, $k_{ON} = 0.25$, $k_{OFF} = 9.75$ and $\beta = 1$.

peaks followed by a gradual decrease due to the degradation.

If we compute the distribution for the mRNA that such a system generates, we obtain a highly asymmetric distribution with a peak around low molecular numbers and a very long right tail produced by bursts. Figure 7.2 shows the mRNA stationary density obtained in two distinct simulations: one via the Gillespie algorithm, corresponding to Figure 7.1, and one from the Poisson-beta structure given in (7.10).

We can visually verify that the two coincide, apart from minor discrepancies due to the finite sample they are based on. Clearly the Poisson-beta has a massive computational advantage and, unlike the former, allows us to obtain an independent sample.

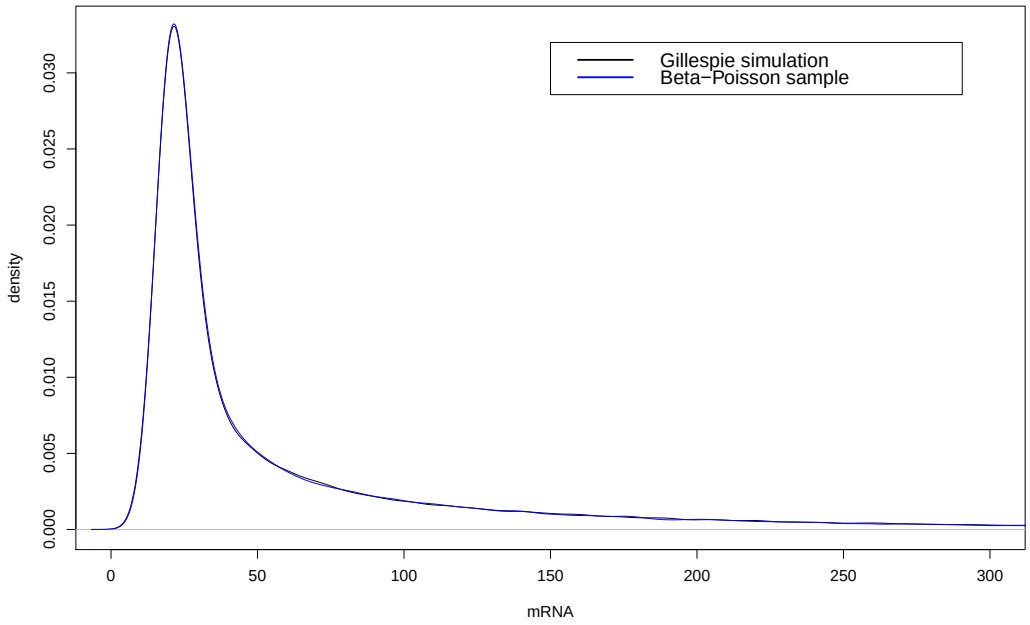


**Figure 7.2:** Marginal density for the mRNA population in a cell at equilibrium obtained both from a Gillespie simulation (in black), for the two-state switch model we propose, and from its Poisson-beta (in blue) stationary distribution. The two densities are almost indistinguishable and minor differences are only due to the limited sample the density is estimated on. Parameters used in the simulation are the same as in Figure 7.1.

## 7.2 Likelihood and latent states

Since every cell contributes to a single observation, we assume that all observations come from cells which have reached their equilibrium and that, within the same experiment, these are independent and identically distributed (iid) with probability function given by (7.1).

We introduce $X = (X_1, ..., X_N)$, the random vector containing the mRNA population from $N$ cells at a single specific time. Given the iid structure, from (7.1), it is straightforward to define a likelihood for a realization of $X$, which we call $x = (x_1, ..., x_N)$, as the product of the densities over the $N$ cells with common parameter vector $\theta = (\tilde{\alpha}_0, \tilde{\alpha}_1, \tilde{k}_{ON}, \tilde{k}_{OFF})$

$$L_X(\theta; x) = \prod_{i=1}^{N} P(X = x_i; \theta). \tag{7.16}$$

As previously described, we do not explicitly compute the density of $X$, and hence we

do not compute $L(x;\theta)$; instead, we will sample $X$ from its stationary distribution.

However, the measurement process is characterized by an additional source of random noise. Observations, which we call $Y = (Y_1, ..., Y_N)$, are assumed to be linked to the original population levels, $X$, via a measurement equation which involves a proportionality constant, $\kappa$, and a normal noise:

$$Y_i = \kappa\, X_i + \epsilon_i, \forall i = 1, ..., N, \tag{7.17}$$

where $\epsilon_i \sim \mathcal{N}(\mu_\epsilon, \sigma_\epsilon^2)$. The mean of the measurement error is non-zero and, as will be shown in the next Chapter, is indeed inferred to be positive. This is due to the fact that the measurement process, in addition to the light intensity of an individual cell, also detects some background signal as part of the cell.

From equation (7.17), we can define the likelihood for the observed data as

$$L_\epsilon(x, \mu_\epsilon, \sigma_\epsilon; y) = \prod_{i=1}^{N} \Phi(y_i - \kappa\, x_i | \mu_\epsilon, \sigma_\epsilon^2), \tag{7.18}$$

where $y = (y_1, ..., y_N)$ represents a realization from $Y$ and $\Phi(.|\mu, \sigma^2)$ denotes the density of a normal r.v. with mean $\mu$ and variance $\sigma^2$.

The measurement equation introduces a latent point for each observation, $X_i$. As illustrated in Figure 7.3, the structure we propose also includes a second level latent state for the probability that the $i$-th gene is ON, which we call $P_i$, where elements for all cells are embedded in $P = (P_1, ..., P_N)$ and are assumed to be independent between different cells and identically distributed. Therefore, we generalise the structure in (7.11), to account for $N$ cells, as

$$X_i \;\sim\; Pois(\tilde{\alpha}_1\, P_i + \tilde{\alpha}_0\, (1 - P_i)), \text{ with } P_i \sim Beta(\tilde{k}_{ON}, \tilde{k}_{OFF}), \text{ for } i = 1, ..., N. \tag{7.19}$$

The analysis is performed in a Bayesian framework where we use a data augmentation procedure (Lesaffre and Lawson, 2012), embedded in a Metropolis-within-Gibbs algorithm, to alternately sample from the conditional distributions of the parameters, given the latent states, and of the latent states, given the parameters.

The details of the algorithm will be described in Section 7.5, after having introduced the hierarchical model and the prior specifications for the parameters.

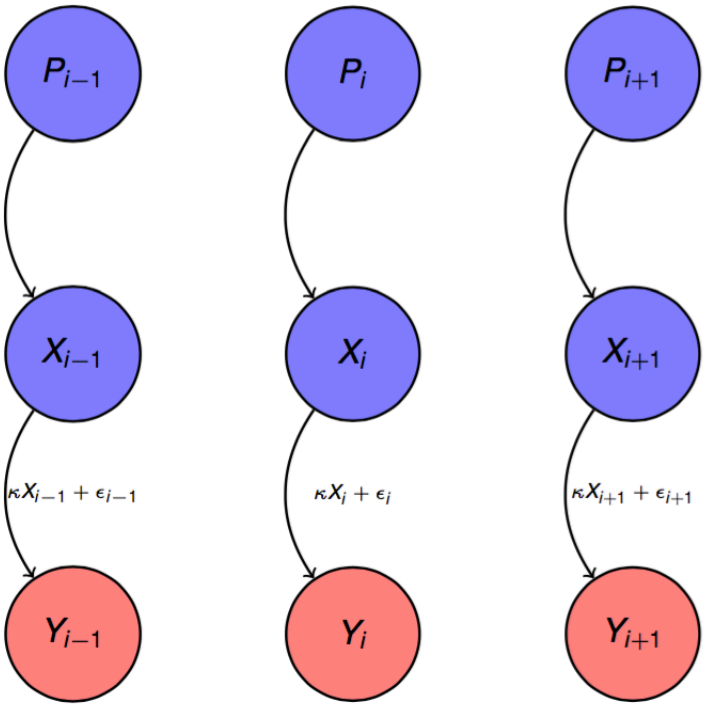


**Figure 7.3:** Two layer latent states, $X_i$ and $P_i$, and observed one, $Y_i$, for $i = 1, ..., N$.

## 7.3 Hierarchical likelihood

All single cell mRNAs from the same experiment are assumed to be iid where, for every experimental condition, four replicates are obtained, each corresponding to approximately 10,000 single cell observations.

In order to pool the information from these replicates and, at the same time, to model the variability between replicates, we embed the model in a hierarchical framework. This allows us to explicitly account for the experimental variability and compare inference, for different experimental conditions, via the hyperparameters.

We define $\mathbf{Y} = (Y^{(1)}, ..., Y^{(K)})$ as the full data, with $Y^{(k)} = (Y_1^{(k)}, ..., Y_{N_k}^{(k)})^T$ representing the data available for the $k$-th experiment, $k = 1, ..., K$, where $N_k$ indicates the number of single cell observations available for experiment $k$. In our case, $K = 4$ and $N_k$ is of the order of approximately 10,000 for each replicate

The hierarchical measurement equation, relating observations $\mathbf{Y}$ to the mRNA populations $\mathbf{X} = (X^{(1)}, ..., X^{(K)})$, with $X^{(k)} = (X_1^{(k)}, ..., X_{N_k}^{(k)})^T$, is

$$Y_i^{(k)} = \kappa^{(k)} X_i^{(k)} + \epsilon_i^{(k)}, \text{ for } i = 1, ..., N_k \text{ and } k = 1, ..., K, \tag{7.20}$$

with $\epsilon_i^{(k)} \sim \mathcal{N}(\mu_\epsilon^{(k)}, \sigma_\epsilon^{2\,(k)})$.

In Figure 7.4 we show the graphical model of this hierarchical system.

Our second layer latent variables are also redefined in the hierarchical context as $\mathbf{P} = (P^{(1)}, ..., P^{(K)})$, with $P^{(k)} = (P_1^{(k)}, ..., P_{N_k}^{(k)})^T$, where $P_i^{(k)}$ is the r.v. representing the probability that gene $i$ of experiment $k$ is in the ON state. We denote by $p^{(k)} = (p_1^{(k)}, ..., p_{N_k}^{(k)})^T$ a realization from $P^{(k)}$. We also define the hierarchical parameter vector for the $k$-th experiment as $\theta^{(k)} = (\tilde{\alpha}_0^{(k)}, \tilde{\alpha}_1^{(k)}, \tilde{k}_{ON}^{(k)}, \tilde{k}_{OFF}^{(k)}, \kappa^{(k)}, \mu_\epsilon^{(k)}, \sigma_\epsilon^{(k)})^T$, including the measurement equation parameters.

Since replicates refer to different cells and are independent, we can define the full likelihood of the data as the product of the likelihoods of the individual replicates:

$$L_\epsilon(\mathbf{X}, \theta; \mathbf{Y}) = \prod_{k=1}^{K} L_\epsilon(x^{(k)}, \theta^{(k)}; y^{(k)}); \tag{7.21}$$

also the density for $\mathbf{X}$ can be defined, following (7.16), as

$$L_X(\theta; \mathbf{X}) = \prod_{k=1}^{K} L_X(\theta^{(k)}; x^{(k)}), \tag{7.22}$$

with $\theta = (\theta^{(1)}, ..., \theta^{(K)})$ and where $L_\epsilon(x^{(k)}, \theta^{(k)}; y^{(k)})$ and $L_X(\theta^{(k)}; x^{(k)})$ represent the density functions, defined in (7.18) and (7.16), for the $k$-th replicate, while $x^{(k)} = (x_1^{(k)}, ..., x_{N_k}^{(k)})^T$ and $y^{(k)} = (y_1^{(k)}, ..., y_{N_k}^{(k)})^T$ indicate realizations from the random elements $X^{(k)}$ and $Y^{(k)}$.

As a notation remark, superscripts between parentheses indicate the replicate, while subscripts refer to a single element of the parameter vector. Each parameter vector, $\theta^{(k)}$, differs in each experimental replicate, yet is assumed to have been generated from a common probability distribution,

$$\theta^{(k)} \sim p(\theta|\Theta)\, \forall k = 1, ..., K,$$

with $\Theta = (\Theta_1, ..., \Theta_p)$, where $p$ is the number of elements in $\theta^{(k)}$, 7 in our case, and $\Theta_j = (\mu_j, \tau_j)^T$ is the hyperparameter vector quantifying the mean and precision of $\theta_j^{(k)}$ over the population of cells, $j = 1, ..., p$. As a prior, we choose a log-normal distribution for each parameter, $p(\theta_j^{(k)}|\Theta) \sim log\mathcal{N}(\mu_j, \frac{1}{\tau_j})$ with $j = 1, ..., p$ and $k = 1, ..., K$, which has $\mathbb{R}^+$ as support, a sensible choice as all parameters are positive.

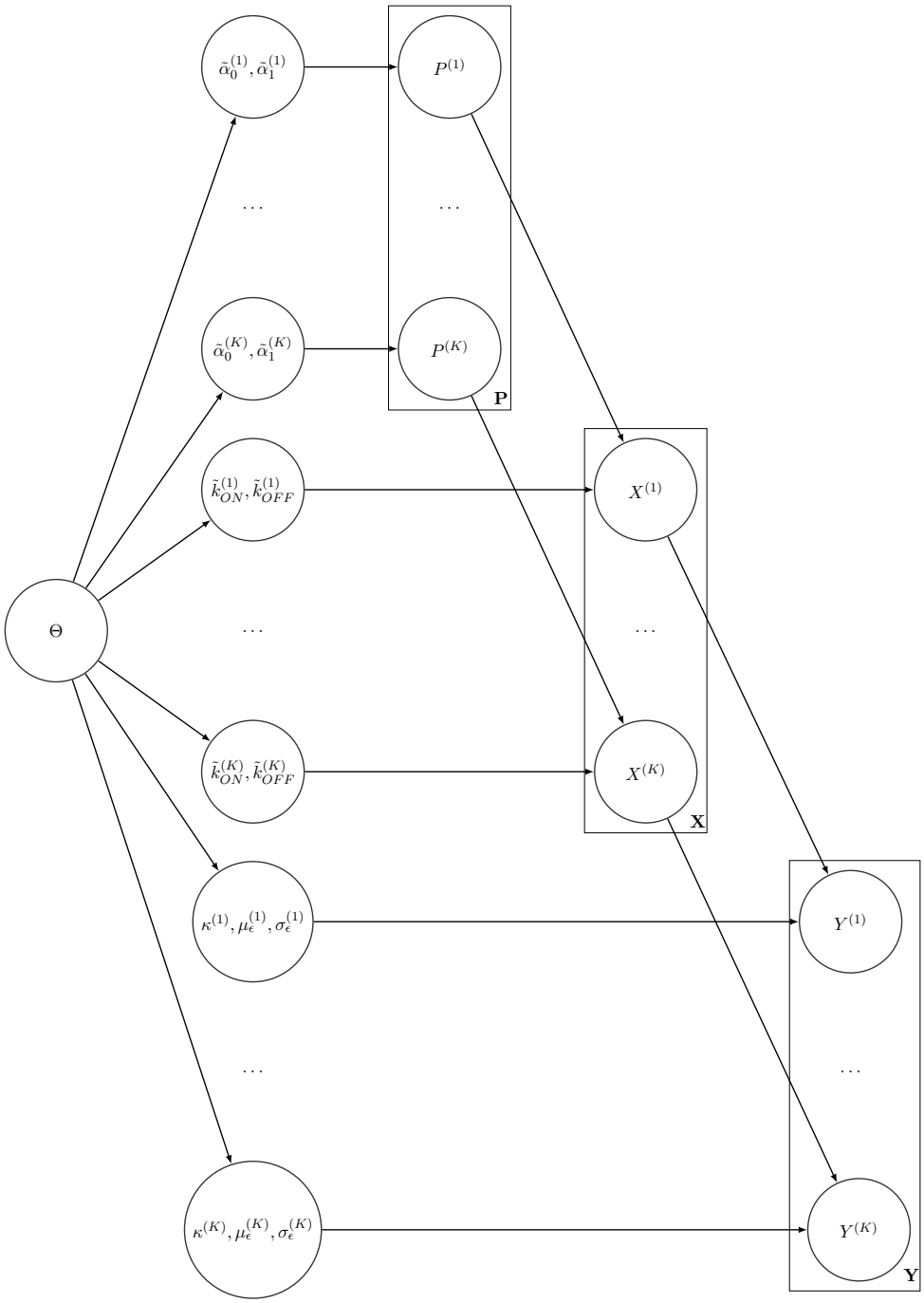


**Figure 7.4:** Graphical model for the hierarchical system. On the left side the hyperparameters $\Theta$ are shown; these generate the hierarchical parameters. From the hierarchical parameters on top the latent states $P^{(1)}, \ldots, P^{(N)}$ are drawn. The latent states $X^{(1)}, \ldots, X^{(N)}$ are then sampled conditional on $P^{(1)}, \ldots, P^{(N)}$ and on the hierarchical parameters in the middle of the graph. Finally, $X^{(1)}, \ldots, X^{(N)}$ together with the hierarchical parameters at the bottom of the graph generate, via the measurement equation, the observed data $Y^{(1)}, \ldots, Y^{(N)}$.

Assuming independence in the prior distributions between parameters and replicates, the prior density of the entire parameter matrix $\theta = (\theta^{(1)}, ..., \theta^{(K)})^T$, conditional on the hyperparameter $\Theta$, is obtained as the double product over the $K$ replicates and the $p$ elements of $\theta^{(k)} = (\theta_1^{(k)}, ..., \theta_p^{(k)})^T$,

$$p(\theta|\Theta) = \prod_{k=1}^{K} p(\theta^{(k)}|\Theta) = \prod_{k=1}^{K}\prod_{j=1}^{p} p(\theta_j^{(k)}|\Theta) = \prod_{k=1}^{K}\prod_{j=1}^{p} p(\theta_j^{(k)}|\Theta_j). \tag{7.23}$$

In the hierarchical framework, we wish to infer the posterior distribution of the hyperparameters given the full data,

$$\begin{aligned} p(\Theta|\mathbf{Y}) &= \int\int p(\theta, \Theta, \mathbf{X}|\mathbf{Y})\, d\theta\, d\mathbf{X} && (7.24)\\ &\propto \int\int L_\epsilon(\mathbf{X}, \theta; \mathbf{Y}) L_X(\theta; \mathbf{X})\, p(\theta|\Theta)\, p(\Theta)\, d\theta\, d\mathbf{X}, && (7.25) \end{aligned}$$

with $p(\Theta)$ being the prior distribution of hyperparameter $\Theta$.

$L_X(\theta; \mathbf{X})$ can be further decomposed as

$$L_X(\theta; \mathbf{X}) \propto \int L_{Pois}(\mathbf{P}, \theta|\mathbf{X})\, L_{Beta}(\theta|\mathbf{P}) d\mathbf{P}, \tag{7.26}$$

where

$$L_{Beta}(\theta|\mathbf{P}) = \prod_{k=1}^{K} L_{Beta}(\theta^{(k)}; p^{(k)}) \tag{7.27}$$

and

$$L_{Pois}(\mathbf{P}, \theta|\mathbf{X}) = \prod_{k=1}^{K} L_{Pois}(p^{(k)}, \theta^{(k)}; x^{(k)}), \tag{7.28}$$

with

$$L_{Beta}(\theta^{(k)}; p^{(k)}) = \prod_{i=1}^{N_k} f_{Beta}(p_i^{(k)}; \tilde{k}_{ON}^{(k)}, \tilde{k}_{OFF}^{(k)}) \tag{7.29}$$

and

$$L_{Pois}(p^{(k)}, \theta^{(k)}; x^{(k)}) = \prod_{i=1}^{N_k} f_{Pois}(x_i^{(k)}|\tilde{\alpha}_0^{(k)} + (\tilde{\alpha}_1^{(k)} - \tilde{\alpha}_0^{(k)})p_i^{(k)}), \tag{7.30}$$

where $f_{Beta}(.|a, b)$ indicates the density of a beta r.v. with parameters $a$ and $b$, i.e. with mean $\dfrac{a}{a+b}$, and $f_{Pois}(.|\lambda)$ indicates the density of a Poisson r.v., with rate $\lambda$, i.e. with mean and variance $\lambda$.

Using equation (7.26), we can rewrite $p(\Theta|\mathbf{Y})$ as

$$
\begin{aligned}
p(\Theta|\mathbf{Y}) &= \int\int p(\theta, \Theta, \mathbf{X}|\mathbf{Y})\, d\theta\, d\mathbf{X} && (7.31)\\
&\propto \int\int L_\epsilon(\mathbf{X}, \theta; \mathbf{Y}) L_X(\theta; \mathbf{X})\, p(\theta|\Theta)\, p(\Theta)\, d\theta\, d\mathbf{X}, && (7.32)\\
&\propto \int\int\int L_\epsilon(\mathbf{X}, \theta; \mathbf{Y})\, L_{Pois}(\mathbf{P}, \theta; \mathbf{X})\, L_{Beta}(\theta; \mathbf{P})\, p(\theta|\Theta)\, p(\Theta)\, d\theta\, d\mathbf{X}\, d\mathbf{P}. && (7.33)
\end{aligned}
$$

## 7.4 Prior and proposal specification

To infer the posterior densities of parameters, we use Markov chain Monte Carlo (MCMC) methods, where the sampling scheme follows a Metropolis-within-Gibbs algorithm (Metropolis and Ulam, 1949; Metropolis *et al.*, 1953; Hastings, 1970).

The prior specification of hyperparameters is analogous to that of the Nrf2 analysis in Part I. The hypermean and hyperprecision of each parameter is assumed to follow a normal, $\mu_j|\tau_j \sim \mathcal{N}(\mu_{j_0}, \frac{\lambda_{j_0}}{\tau_j})$, and a gamma distribution, $\tau_j \sim \mathcal{G}(a_{j_0}, b_{j_0})$, respectively, where $\mathcal{G}(a_{j_0}, b_{j_0})$ indicates the gamma r.v. parametrized with mean $\frac{a_{j_0}}{b_{j_0}}$ and variance $\frac{a_{j_0}}{b_{j_0}^2}$. This structure is conjugate with the log-normal prior and thus we can sample the hyperparameters via a Gibbs step. Excluding $\mu_\epsilon$ and $\sigma_\epsilon$, for which an informative prior is formulated, the hyperpriors are chosen to be non-informative for all the remaining parameters, with $\mu_{j_0} = 0$, $\lambda = 10^4$, $a_{j_0} = 0.001$ and $b_{j_0} = 0.001$, which is a standard choice (Gelman, 2006) which we also used in the Nrf2 analysis in Part I.

### Informative prior for $\mu_\epsilon$ and $\sigma_\epsilon$

Following exploratory data analyses on the background noise in all four replicates, which will be shown in Section 8.1, we formulate an informative prior on the measurement error parameters of each replicate separately. In particular, we use a constant prior for $\mu_\epsilon$ and $\sigma_\epsilon$, with different hyperparameters for each experiment, therefore the hypermean and hyperprecision parameters for $\mu_\epsilon$ and $\sigma_\epsilon$ are not updated. In the log space, the prior is formulated as

$$log(\mu_\epsilon^{(k)}) \sim \mathcal{N}(\mu_{\mu_\epsilon}^{(k)}, \sigma_{\mu_\epsilon}^{2\,(k)}) \text{ and } log(\sigma_\epsilon^{(k)}) \sim \mathcal{N}(\mu_{\sigma_\epsilon}^{(k)}, \sigma_{\sigma_\epsilon}^{2\,(k)}),\ k = 1, ..., 4,$$

with

$$
\begin{aligned}
(\mu^{(1)}_{\mu_\epsilon}, \mu^{(2)}_{\mu_\epsilon}, \mu^{(3)}_{\mu_\epsilon}, \mu^{(4)}_{\mu_\epsilon}) &= (6.65, 6.69, 6.87, 6.79), && (7.34)\\
(\mu^{(1)}_{\sigma_\epsilon}, \mu^{(2)}_{\sigma_\epsilon}, \mu^{(3)}_{\sigma_\epsilon}, \mu^{(4)}_{\sigma_\epsilon}) &= (5.80, 5.74, 5.98, 5.72), && (7.35)\\
(\sigma^{(1)}_{\mu_\epsilon}, \sigma^{(2)}_{\mu_\epsilon}, \sigma^{(3)}_{\mu_\epsilon}, \sigma^{(4)}_{\mu_\epsilon}) &= 10^{-3}\,(5.54, 6.45, 5.34, 3.37), && (7.36)\\
(\sigma^{(1)}_{\sigma_\epsilon}, \sigma^{(2)}_{\sigma_\epsilon}, \sigma^{(3)}_{\sigma_\epsilon}, \sigma^{(4)}_{\sigma_\epsilon}) &= 10^{-3}\,(8.80, 11.85, 9.18, 6.91). && (7.37)
\end{aligned}
$$

### Adaptive random walk proposal

The sampling of the hierarchical parameters in $\theta^{(k)}$ follows an analogous approach to the study in Part I. After an initial analysis, where each hierarchical parameter is proposed independently from a simple random walk (RW), we study the correlation between the posterior chains and merge, in the same block, the most correlated parameters, and obtain 3 blocks, which we call $\theta^{(k)}_{(b_1)} = (\tilde{\alpha}^{(k)}_1, \tilde{\alpha}^{(k)}_0, \kappa^{(k)})$, $\theta^{(k)}_{(b_2)} = (\tilde{k}^{(k)}_{ON}, \tilde{k}^{(k)}_{OFF})$ and $\theta^{(k)}_{(b_3)} = (\mu^{(k)}_\epsilon, \sigma^{(k)}_\epsilon)$. We also define $b_1 = \{1, 2, 5\}$, $b_2 = \{3, 4\}$ and $b_3 = \{6, 7\}$ as the vectors containing the elements of $\theta^{(k)}$ belonging to the first, second and third block, respectively.

Values in each block are proposed from an adaptive random walk (ARW) scheme (Haario *et al.*, 2001) where, in the implementation from Haario *et al.* (2001), constants $\epsilon$ and $s_d$ are tuned in order to optimize each block's acceptance rate. Each block is proposed, in the log space, from a normal distribution, centred around the previous iteration values, with variance proportional to the covariance matrix estimated from the parameter chains of the respective block.

The algorithm is first run for 2,000 iterations without adaptation, as a standard random walk (RW). The covariance matrices are computed on each chains, excluding the first 1,000 values, and then used in the proposal variance. Again, as we have noted in Part I, the diminishing adaptation requirement (Roberts and Rosenthal, 2009) is respected; the influence that the parameter values have, at the $r$-th iteration of the MCMC, on the proposal variances, goes to 0 as $r \to \infty$.

## 7.5 Metropolis-within-Gibbs algorithm

In this Section we illustrate the algorithm used to alternately sample from the posterior densities of parameters and latent states. Given all hierarchical parameters in $\theta^{(k)}$ are positive, we sample them in the log space; the prior for $log(\theta^{(k)}_j)$ becomes $p(log(\theta^{(k)}_j)|\Theta) \sim \mathcal{N}(\mu_j, \frac{1}{\tau_j})$ for each $k = 1, ..., 4$ and $j = 1, ..., 7$. We implement inference via a Metropolis-within-Gibbs algorithm (Metropolis and Ulam, 1949; Metropo-

lis *et al.*, 1953; Hastings, 1970). The key idea of our data augmentation scheme is to separate, in each replicate $k$, the sampling of the latent states, $X^{(k)}$ and $P^{(k)}$, and to make use of the conditional densities $L_{Beta}(\theta^{(k)}; p^{(k)})$ and $L_{Pois}(p^{(k)}, \theta^{(k)}; x^{(k)})$ to sample the hierarchical parameters $\theta^{(k)}$ without the need to compute the density of $X^{(k)}$ explicitly, as in (7.16), which would involve the unstable and complex computation of ${}_1F_1$.

We now give the structure of the sampling scheme in detail.

**Initialisation**: We initialize $\Theta$ and $\theta$ to some arbitrary positive real values. In our real data analysis, initial values are chosen after some exploratory data studies. For each replicate $k = 1, ..., 4$, we initialize $X^{(k)} = \dfrac{Y^{(k)}}{20}$, a plausible choice after preliminary analyses, and $P_i^{(k)}$ by randomly drawing from a Poisson distribution $Pois(\tilde{k}_{ON}^{(k)}, \tilde{k}_{OFF}^{(k)})$, $\forall i = 1, ..., N_k$, for chosen starting values of $\tilde{k}_{ON}^{(k)}$ and $\tilde{k}_{OFF}^{(k)}$.

**Update**: At each iteration of the MCMC $r = 1, ..., R$, where $R$ denotes the total number iterations, we update parameters by repeating the following steps.

Update $\Theta|\theta$**:** Each hyper-parameter $\Theta_j = (\mu_j, \tau_j)^T$, $j = 1, ..., 7$, is sampled, conditional on the corresponding hierarchical parameters $\theta_j = (\theta_j^{(1)}, ..., \theta_j^{(4)})$, from a Gibbs step owing to the hyperprior conjugate choice. In spite of using a constant prior for the measurement error parameters, we still sample the respective hyperprior parameters, although not being used in the sampling of the hierarchical ones. The hypermean, *a posteriori*, still is normal,

$$\mu_j|\tau_j \sim \mathcal{N}\left(\frac{\tau_j/\lambda_{j_0}\,\mu_{j_0} + \tau_j \sum_{k=1}^{4} log(\theta_j^{(k)})}{\tau_j/\lambda_{j_0} + 4\,\tau_j}, (\tau_j/\lambda_{j_0} + 4\,\tau_j)^{-1}\right),$$

with $\mathcal{N}(\mu, \sigma^2)$ being the univariate normal distribution with mean $\mu$ and variance $\sigma^2$, and the hyperprecision, *a posteriori*, is again gamma distributed (Gelman, 2006),

$$\tau_j|\mu_j \sim Gamma\left(\alpha_{j_0} + 4/2, \beta_{j_0} + 1/2 \sum_{k=1}^{4} \left(log(\theta_j^{(k)}) - \mu_j\right)^2\right),$$

where $Gamma(a, b)$ denotes the gamma r.v. with shape and rate parameters $a$ and $b$, respectively, i.e. with mean $\dfrac{a}{b}$ and variance $\dfrac{a}{b^2}$.

Update $\theta|\Theta, \mathbf{Y}, \mathbf{X}, \mathbf{P}$: For every replicate, $k = 1, ..., 4$, the parameters in each block, $j = 1, 2, 3$, are sampled independently in the log space, via a Metropolis step, where $log(\theta_{(b_j)}^{*(k)})$ is proposed from $\mathcal{N}\left(log(\theta_{(b_j)}^{(k)}), \Sigma_{ARW_j}^{(k)}\right)$, with $\Sigma_{ARW_j}^{(k)}$ being the

ARW covariance matrix, for block $j$ and replicate $k$, and where $\theta^{(k)}_{(b_j)}$ and $\theta^{*(k)}_{(b_j)}$ denote the current iteration and the proposed parameter values for block $j$, respectively.

Acceptance rates for the three blocks are computed as follows:

Block $b_1$: The proposal for the first block, $\theta^{*(k)}_{(b_1)} = (\tilde{\alpha}^{*(k)}_1, \tilde{\alpha}^{*(k)}_0, \kappa^{*(k)})$, is accepted with probability $min(1, A^{(k)}_1)$, with

$$A^{(k)}_1 = \frac{L_{Pois}\left(p^{(k)}, \tilde{\theta}^{*(k)}_{\{b_1\}}; x^{(k)}\right) L_\epsilon\left(x^{(k)}, \tilde{\theta}^{*(k)}_{\{b_1\}}; y^{(k)}\right) \prod_{j\in b_1} \Phi(log(\theta^{*(k)}_j); \mu_j, \frac{1}{\tau_j})}{L_{Pois}(p^{(k)}, \theta^{(k)}; x^{(k)})\, L_\epsilon\left(x^{(k)}, \theta^{(k)}; y^{(k)}\right) \prod_{j\in b_1} \Phi(log(\theta^{(k)}_j); \mu_j, \frac{1}{\tau_j})},$$

where $\tilde{\theta}^{*(k)}_{\{b_1\}} = \left(\tilde{\alpha}^{*(k)}_0, \tilde{\alpha}^{*(k)}_1, \tilde{k}^{(k)}_{ON}, \tilde{k}^{(k)}_{OFF}, \kappa^{*(k)}, \mu^{(k)}_\epsilon, \sigma^{(k)}_\epsilon\right)^T$ is the vector with elements in block $b_1$ from the proposed values $\theta^{*(k)}_{(b_1)}$ and with parameters for blocks $b_2$ and $b_3$ from the current iteration values $\theta^{(k)}_{(b_2)}$ and $\theta^{(k)}_{(b_3)}$, while functions $L_{Pois}$ and $L_\epsilon$ have been specified in (7.30) and (7.18), respectively.

Block $b_2$: The new candidate for the second block, $\theta^{*(k)}_{(b_2)} = (\tilde{k}^{*(k)}_{ON}, \tilde{k}^{*(k)}_{OFF})$, is accepted with probability $min(1, A^{(k)}_2)$, with

$$A^{(k)}_2 = \frac{L_{Beta}\left(\tilde{\theta}^{*(k)}_{\{b_2\}}; p^{(k)}\right) \prod_{j\in b_2} \Phi(log(\theta^{*(k)}_j); \mu_j, \frac{1}{\tau_j})}{L_{Beta}\left(\theta^{(k)}; p^{(k)}\right) \prod_{j\in b_2} \Phi(log(\theta^{(k)}_j); \mu_j, \frac{1}{\tau_j})},$$

where, analogous to above, $\tilde{\theta}^{*(k)}_{\{b_2\}} = \left(\tilde{\alpha}^{(k)}_0, \tilde{\alpha}^{(k)}_1, \tilde{k}^{*(k)}_{ON}, \tilde{k}^{*(k)}_{OFF}, \kappa^{(k)}, \mu^{(k)}_\epsilon, \sigma^{(k)}_\epsilon\right)^T$ and $L_{Beta}$ is the function in (7.29).

Block $b_3$: The proposal for the third block, $\tilde{\theta}^{(k)}_{(b_3)} = (\mu^{*(k)}_\epsilon, \sigma^{*(k)}_\epsilon)$, is accepted with probability $min(1, A^{(k)}_3)$, with

$$A^{(k)}_3 = \frac{L_\epsilon\left(x^{(k)}, \tilde{\theta}^{*(k)}_{\{b_3\}}; y^{(k)}\right) \Phi(log(\mu^{*(k)}_\epsilon); \mu^{(k)}_{\mu_\epsilon}, \sigma^{2\,(k)}_{\mu_\epsilon})\, \Phi(log(\sigma^{*(k)}_\epsilon); \mu^{(k)}_{\sigma_\epsilon}, \sigma^{2\,(k)}_{\sigma_\epsilon})}{L_\epsilon\left(x^{(k)}, \theta^{(k)}; y^{(k)}\right) \Phi(log(\mu^{(k)}_\epsilon); \mu^{(k)}_{\mu_\epsilon}, \sigma^{2\,(k)}_{\mu_\epsilon})\, \Phi(log(\sigma^{(k)}_\epsilon); \mu^{(k)}_{\sigma_\epsilon}, \sigma^{2\,(k)}_{\sigma_\epsilon})},$$

with $\tilde{\theta}^{*(k)}_{\{b_3\}} = \left(\tilde{\alpha}^{(k)}_0, \tilde{\alpha}^{(k)}_1, \tilde{k}^{(k)}_{ON}, \tilde{k}^{(k)}_{OFF}, \kappa^{(k)}, \mu^{*(k)}_\epsilon, \sigma^{*(k)}_\epsilon\right)^T$, where the constant priors have been specified in the previous Section.

Update $\mathbf{P}|\theta, \mathbf{X}$: For every replicate $k = 1, ..., 4$, each second level latent state, $P^{(k)}_i$, is sampled independently, for $i = 1, ..., N_k$, from its conditional density,

$$f_4(p^{(k)}_i|\theta^{(k)}, X^{(k)}_i = x^{(k)}_i) = f_{Beta}(p^{(k)}_i|\tilde{k}^{(k)}_{ON}, \tilde{k}^{(k)}_{OFF})\, f_{Pois}(x^{(k)}_i|\tilde{\alpha}^{(k)}_0 + (\tilde{\alpha}^{(k)}_1 - \tilde{\alpha}^{(k)}_0)p^{(k)}_i).$$

In order to increase the acceptance rate of the Metropolis-Hastings step, the candidate for $P_i^{(k)}$, which we call $\tilde{p}_i^{(k)}$, is proposed from $f_{Beta}(.|\tilde{k}_{ON}^{(k)}, \tilde{k}_{OFF}^{(k)})$. The acceptance rate for $\tilde{p}_i^{(k)}$ is $min(1, A_P^{(k)})$, where $A_P^{(k)}$ simplifies to

$$A_P^{(k)} = \frac{f_{Pois}(x_i^{(k)}|\tilde{\alpha}_0^{(k)} + (\tilde{\alpha}_1^{(k)} - \tilde{\alpha}_0^{(k)})\tilde{p}_i^{(k)}).}{f_{Pois}(x_i^{(k)}|\tilde{\alpha}_0^{(k)} + (\tilde{\alpha}_1^{(k)} - \tilde{\alpha}_0^{(k)})p_i^{(k)}).}.$$

Update $\mathbf{X}|\theta, \mathbf{P}, \mathbf{Y}$: In each replicate $k = 1, ..., 4$, every latent state for the mRNA population, $X_i^{(k)}$, is sampled independently, for $i = 1, ..., N_k$, from the following conditional density

$$\begin{aligned} f_5(x_i^{(k)}|\theta^{(k)}, P_i^{(k)} = p_i^{(k)}, Y_i^{(k)} = y_i^{(k)}) &= f_{Pois}(x_i^{(k)}|\tilde{\alpha}_0^{(k)} + (\tilde{\alpha}_1^{(k)} - \tilde{\alpha}_0^{(k)})p_i^{(k)}) \times \\ &\quad \Phi(y_i^{(k)} - \kappa^{(k)}\, x_i^{(k)}|\mu_\epsilon^{(k)}, \sigma_\epsilon^{2(k)}). \end{aligned}$$

Similar to the updating step for $P_i$, also $X_i$ is proposed in such a way as to increase the acceptance rate of the Metropolis-Hastings step. In particular, the proposal for $X_i^{(k)}$, which we denote by $\tilde{x}_i^{(k)}$, is sampled from $f_{Pois}(.|\tilde{\alpha}_0^{(k)} + (\tilde{\alpha}_1^{(k)} - \tilde{\alpha}_0^{(k)})p_i^{(k)})$.
The acceptance rate for $\tilde{x}_i^{(k)}$ is $min(1, A_X^{(k)})$, where $A_X^{(k)}$ simplifies to

$$A_X^{(k)} = \frac{\Phi(y_i^{(k)} - \kappa^{(k)}\, \tilde{x}_i^{(k)}|\mu_\epsilon^{(k)}, \sigma_\epsilon^{2(k)}).}{\Phi(y_i^{(k)} - \kappa^{(k)}\, x_i^{(k)}|\mu_\epsilon^{(k)}, \sigma_\epsilon^{2(k)})}.$$

## 7.6 Simulation study

In order to assess the validity of the methodology described above, we carry out a simulation study. We simulate three hierarchical databases with an analogous structure to the observed ones, composed of 4 replicates of 10,000 independent observations each, a similar size to the available experimental data.

In each simulation study, we initially choose a set of hyperprior values $\Theta$, and sample 4 independent parameter vectors $\theta^{(k)} \sim p(.|\Theta)$, $k = 1, ..., 4$. In particular, each element $\theta_j^{(k)}$ is sampled from a $log\mathcal{N}(\mu_j, \frac{1}{\tau_j})$, for $k = 1, ..., 4$ and $j = 1, ...5$. The last two elements of $\theta^{(k)}$, i.e. the measurement mean and standard deviation, are fixed to their prior mode, i.e. $\mu_\epsilon^{(k)} = e^{\mu_{\mu_\epsilon}^{(k)}}$ and $\sigma_\epsilon^{(k)} = e^{\mu_{\sigma_\epsilon}^{(k)}}$, for $k = 1, ..., 4$, where $\mu_{\mu_\epsilon}^{(k)}$ and $\mu_{\sigma_\epsilon}^{(k)}$ have been defined in Section 7.4.

For each parameter $\theta^{(k)} = (\tilde{\alpha}_0^{(k)}, \tilde{\alpha}_1^{(k)}, \tilde{k}_{ON}^{(k)}, \tilde{k}_{OFF}^{(k)}, \kappa^{(k)}, \mu_\epsilon^{(k)}, \sigma_\epsilon^{(k)})^T$, and independently for every $i = 1, ..., N_k = 10^4$, we sample, $P_i^{(k)} \sim Beta(\tilde{k}_{ON}^{(k)}, \tilde{k}_{OFF}^{(k)})$ and, in turns, $X_i^{(k)} \sim Pois((\tilde{\alpha}_1^{(k)} - \tilde{\alpha}_0^{(k)})\, P_i^{(k)} + \tilde{\alpha}_0^{(k)})$. We then sample the measurement

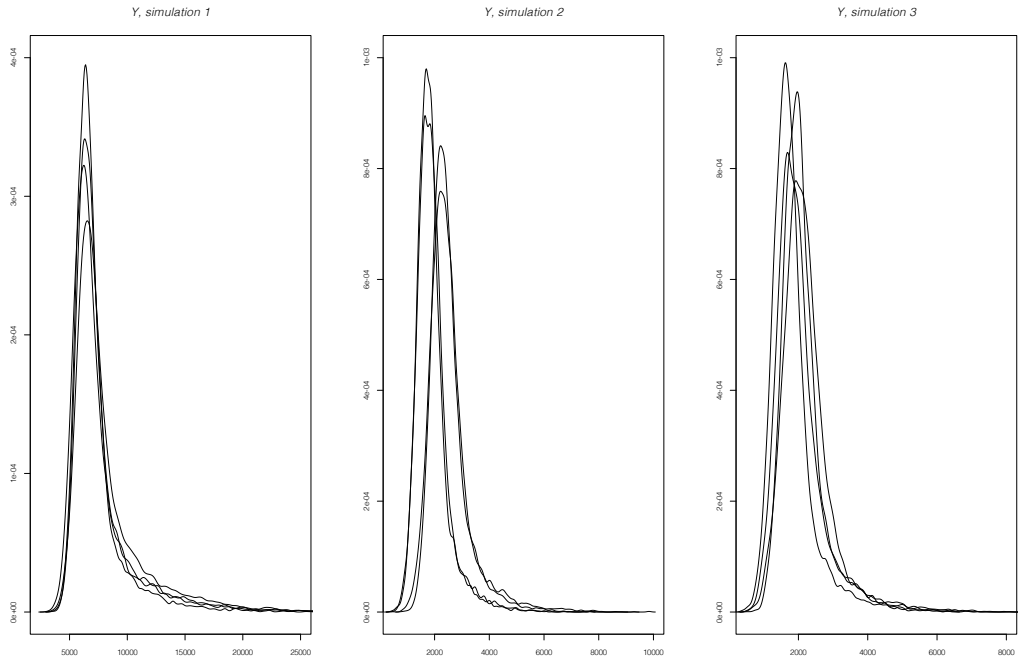


**Figure 7.5:** Densities of the simulated data from three simulation studies. Each image refers to the four replicates of a simulation.

error, $\epsilon_i^{(k)} \sim \mathcal{N}(\mu_\epsilon^{(k)}, \sigma_\epsilon^{2\,(k)})$, and obtain the simulated observed value for $Y_i^{(k)}$ as $Y_i^{(k)} = \kappa^{(k)}\, X_i^{(k)} + \epsilon_i^{(k)}$, $i = 1, ..., N_k$. By repeating this scheme on the 4 replicates, we obtain the simulated hierarchical data embedded in $\mathbf{Y} = (Y^{(1)}, Y^{(2)}, Y^{(3)}, Y^{(4)})$.

The procedure is repeated on each one of the three simulations to obtain three such simulated datasets. hyperparameters are chosen in order to reproduce approximatively similar densities to the ones observed for the experimental data. All sampled hierarchical parameter values are reported in Tables A.5-A.7 in the Appendix. The densities for the simulated data, $Y^{(k)}$, are shown in Figures 7.5. We also display the densities for the latent states in $P^{(k)}$ and $X^{(k)}$ in Figures B.74 and B.75, respectively.

For each simulated database, $\mathbf{Y}$, we perform inference, as described in Sections 7.5, to check the performance of the MCMC algorithm and to very that our methodology allows us to retrieve the original hierarchical parameter values used to simulate the data.

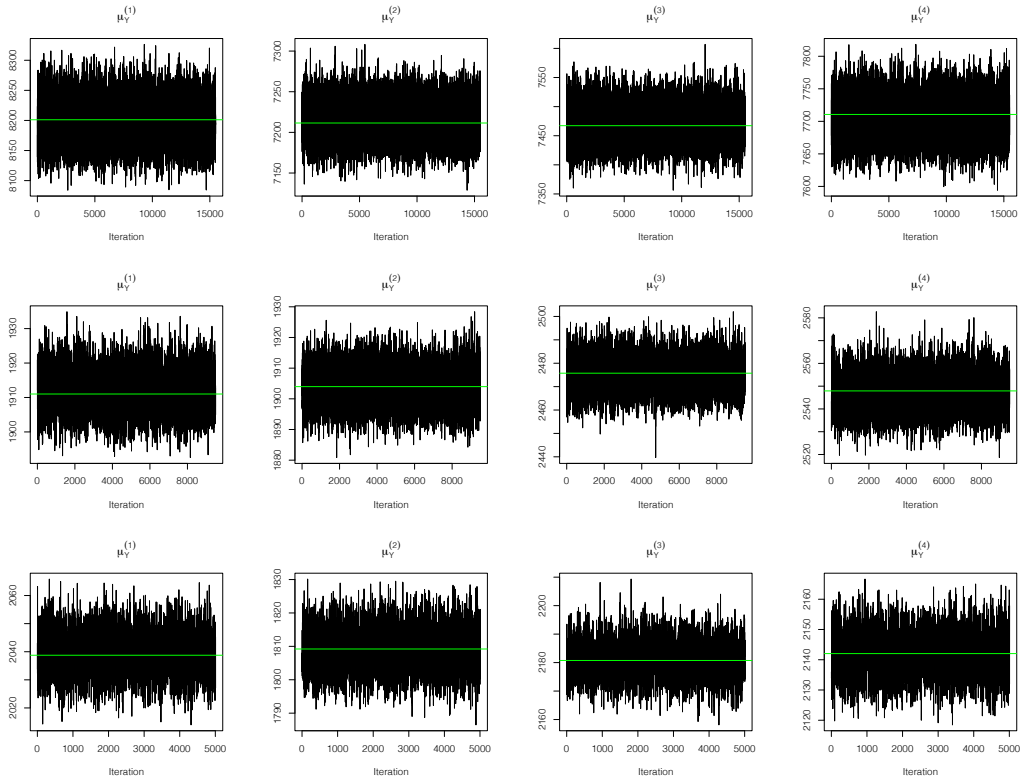


**Figure 7.6:** Trace plots for the hierarchical posterior chains of $\mu_Y^{(k)}$, for all experiments. Rows, from top to bottom, correspond to simulation studies 1, 2 and 3, respectively. Columns, from left to right, refer to the the simulated experiments 1 to 4. The horizontal green lines denote the mean of the observations in the respective experiment. Chains are thinned with tinning factor equals to 100.

## Inferential Results

We perform inference, on each simulated hierarchical dataset, as described in Sections 7.5. The initial iterations of the MCMC are discarded as *burn-in*; after reaching convergence, the chains are run for at least $5 \times 10^5$ more iterations. The acceptance rates for blocks $b_1$, $b_2$ and $b_3$ are 0.04, 0.26 and 0.68, in the first simulation, 0.04, 0.15 and 0.52, in the second simulation, and 0.03, 0.12 and 0.52, in the third simulation. The average acceptance rates for each point of the latent states, $P_i^{(k)}$ and $X_i^{(k)}$, are 0.52 and 0.34, in the first simulation, 0.57 and 0.78, in the second simulation, and 0.52 and 0.80, in the third simulation.

Inference is successful on both hyper and hierarchical parameters. Figure 7.6 shows the trace plots, for each of the three simulations, of the chains for the mean of the observed data of each experiment, i.e. $\mu_Y^{(k)} = E(Y_i^{(k)})$, for $k = 1, ..., 4$ and

$i = 1, ..., N_k$. Following the measurement equation in (7.20), this quantity is defined as

$$\mu_Y^{(k)} = E(Y_i^{(k)}) = \kappa^{(k)}\, E(X_i^{(k)}) + \mu_\epsilon^{(k)}, \tag{7.38}$$

where $E(X_i^{(k)})$ is specified in (7.14). Figures B.76-B.78 in the Appendix display the posterior densities for the hypermean parameters of each simulation, with the vertical line denoting the real value used in the simulation. We can see that the original values of the hypermean parameters are always in a central and high density area of the corresponding posterior distribution.

Table 7.1 reports the coverages of the high posterior density (HPD) credible intervals (CIs) for the hierarchical parameters. CIs are computed for confidence levels 0.9 and 0.95; in our simulation study both have the same coverage: on average 94% of parameters fall in both 0.90 and 0.95 level HPD CIs. Concerning the hyperparameters, all 15 hypermean and hyperprecision parameters, 5 for each experiment, fall in the respective 0.90 level HPD CI.

| Count | $\tilde{\alpha}_0^{(k)}$ | $\tilde{\alpha}_1^{(k)}$ | $\tilde{k}_{ON}^{(k)}$ | $\tilde{k}_{OFF}^{(k)}$ | $\kappa^{(k)}$ | $\mu_\epsilon^{(k)}$ | $\sigma_\epsilon^{(k)}$ | Tot. |
|---|---|---|---|---|---|---|---|---|
| Out of 12 | 12 | 12 | 10 | 11 | 10 | 12 | 12 | 79/84 |
| % | 100.0 | 100.0 | 83.3 | 91.7 | 83.3 | 100.0 | 100.0 | 94.0 |

**Table 7.1:** Coverages of, both 0.90 and 0.95 level HPD CIs for the 7 hierarchical parameters in the three simulation studies in the four experiments. Column "Tot." represents the overall coverage over the 7 hierarchical parameters. Since each of the three simulations has 4 simulated replicates, the total number of hierarchical elements for each parameter adds to 12.

We also study the mean and standard deviation of the latent states in each replicate, $k = 1, ..., 4$, namely $\mu_P^{(k)} = E(P_i^{(k)})$, $\sigma_P^{(k)} = \sqrt{Var(P_i^{(k)})}$, $\mu_X^{(k)} = E(X_i^{(k)})$ and $\sigma_X^{(k)} = \sqrt{Var(X_i^{(k)})}$, with $i = 1, ..., N_k$. These elements have been defined in (7.12)-(7.15); we now introduce their hierarchical counterparts for the $k$-th experiment, $k = 1, ..., 4$, as:

$$\mu_P^{(k)} = \frac{\tilde{k}_{ON}^{(k)}}{\tilde{k}_{ON}^{(k)} + \tilde{k}_{OFF}^{(k)}} \tag{7.39}$$

$$\sigma_P^{2\,(k)} = \frac{\tilde{k}_{ON}^{(k)}\, \tilde{k}_{OFF}^{(k)}}{(\tilde{k}_{ON}^{(k)} + \tilde{k}_{OFF}^{(k)})^2\, (\tilde{k}_{ON}^{(k)} + \tilde{k}_{OFF}^{(k)} + 1)} \tag{7.40}$$

$$\mu_X^{(k)} = \tilde{\alpha}_0^{(k)} + (\tilde{\alpha}_1^{(k)} - \tilde{\alpha}_0^{(k)})\, \frac{\tilde{k}_{ON}^{(k)}}{\tilde{k}_{ON}^{(k)} + \tilde{k}_{OFF}^{(k)}} \tag{7.41}$$

$$\sigma_X^{2\,(k)} = E(X^{(k)}) + \frac{\tilde{k}_{ON}^{(k)}\, \tilde{k}_{OFF}^{(k)}\, (\tilde{\alpha}_1^{(k)} - \tilde{\alpha}_0^{(k)})^2}{(\tilde{k}_{ON}^{(k)} + \tilde{k}_{OFF}^{(k)})^2\, (\tilde{k}_{ON}^{(k)} + \tilde{k}_{OFF}^{(k)} + 1)}. \tag{7.42}$$

Table 7.2 reports the coverages for these parameters. Results are satisfactory, particularly for the mRNA population, with a 96% average coverage of the 0.90 and 0.95 level HPD CIs.

| Count | $\mu_P^{(k)}$ | $\sigma_P^{(k)}$ | $\mu_X^{(k)}$ | $\sigma_X^{(k)}$ | Tot. |
|---|---|---|---|---|---|
| Out of 12 | 11 | 11 | 12 | 12 | 46/48 |
| % | 91.7 | 91.7 | 100.0 | 100.0 | 95.8 |

**Table 7.2:** Coverages of, both, 0.90 and 0.95 level HPD CIs for the latent states means and standard deviations in the three simulation studies. Column "Tot." represents the overall coverage over the four hierarchical parameters considered.

## 7.7 Summary

In this Chapter we proved that, in a two-state switch gene model with transcription in both states, the stationary distribution of the mRNA population in single cells corresponds to a Poisson-beta. We also introduced a measurement equation relating the observations to the original mRNA counts, which are treated as latent states. Indeed, we use a two layer latent structure, with the second layer modelling the probability that the gene is in the ON state. All observations from the same experiment are assumed to be iid; nevertheless we assume a hierarchy on the four experimental replicates in our data. We define the likelihood of such a model and use Bayesian hierarchical methods, via MCMC methods, to infer the parameters' posterior chains. We illustrated prior choices, involving informative fixed priors for the measurement error mean and standard deviation. We then described our Metropolis-within-Gibbs algorithm, which involves the conditional sampling of parameters and latent states, via a data augmentation procedure. Finally, we tested our methodology on three independent simulation studies and showed that it is able to retrieve the original parameter values.

In the next Chapter we will initially describe an exploratory data analysis on background noise data which will help us to formulate a constant informative prior for the measurement error mean and standard deviation in each replicate. We will then apply the methodology described in this Chapter to our experimental data and compare results across experimental conditions.

# Chapter 8

# Inference on experimental data

> I really don't trust statistics much. A man with his head in a hot oven and his feet in a freezer has statistically an average body temperature.
>
> *Charles Bukowski*

This Chapter focuses on the hierarchical analysis of the available data. We first describe an exploratory analysis of the background noise data, which allows us to formulate informative priors for the measurement error parameters. We then apply the methodology described in Section 7.5 to the experimental data and compare results under the different experimental conditions.

| *Tetracycline* | Experiment | | | | |
|---|---|---|---|---|---|
| ng/ml | 1 | 2 | 3 | 4 | Average |
| 0 | 6,041 | 7,441 | 14,129 | 11,057 | 9,667.00 |
| 5 | 5,490 | 8,267 | 13,803 | 11,121 | 9,670.25 |
| 10 | 5,621 | 6,011 | 13,955 | 10,558 | 9,036.25 |
| Background | 6,257 | 3,536 | 5,487 | 10,891 | 6,542.75 |

**Table 8.1:** Number of available observations for the experimental data for each of the four replicates, under various level of induction by *tetracycline*, expressed in nanograms per millilitre (ng/ml). "Background" refers to the background data, used in Section 8.1 to infer the measurement error parameters.

The number of observations under each condition is reported in Table 8.1, while the observed densities of the number mRNA molecules in single cells are plotted in Figures 8.1. It is evident how the densities are more peaked around lower values, when not induced, and gradually move towards higher values with a longer

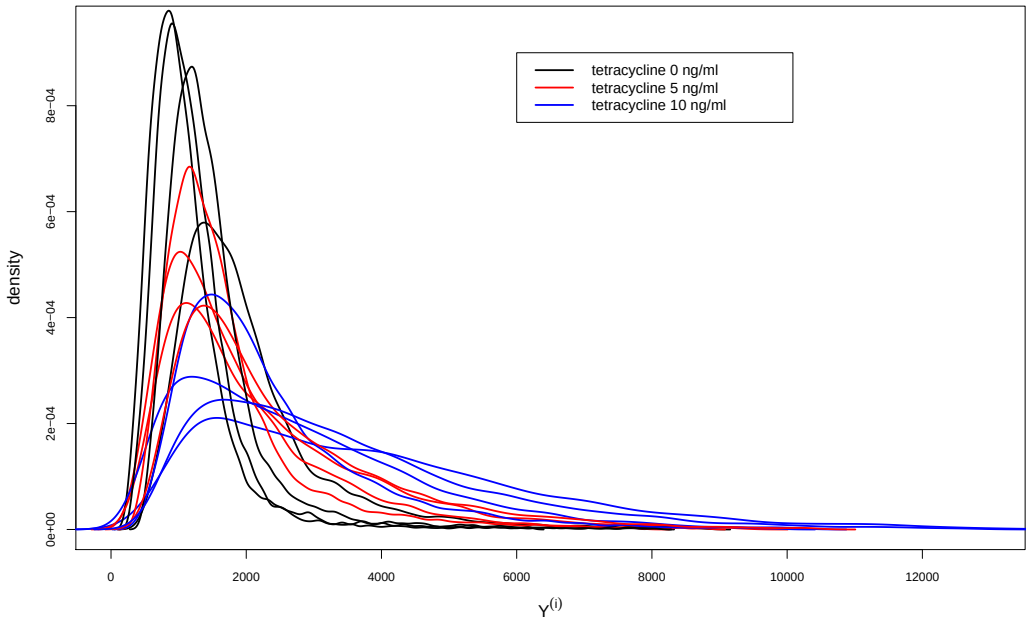


**Figure 8.1:** Empirical densities of the number of mRNA molecules in single cells computed on the experimental data, for the fours replicates, under the induction of *tetracycline* with 0 (in black), 5 (in red) and 10 (in blue) ng/ml.

right tail, as the stimulation level increases. Table A.8 reports the average value of observations in each experiment, $\langle Y^{(k)} \rangle$, $k = 1, ..., 4$. We can see how this value increases with the level of stimulation and, on average across the four experiments, it approximately doubles when going from from 0 to 10 ng/ml of *tetracycline*. It is also clear that densities referring to different replicates present non-negligible variations, hence justifying the use of a hierarchical approach between replicates.

## 8.1 Exploratory analysis of the measurement error

This Section focuses on a study of the measurement error, which will allow us to formulate an informative prior for the error mean and standard deviation parameters in each replicate, $\mu_\epsilon^{(k)}$ and $\sigma_\epsilon^{(k)}$, $k = 1, ..., 4$. Since all the experimental conditions belonging to same replicate are measured in the same experiment, it is reasonable to assume that, *a priori*, that they share the same parameters for the measurement error distribution. Indeed, *a posteriori*, we allow all hierarchical parameters to vary in each experimental condition and replicate, yet we formulate a distinct informative

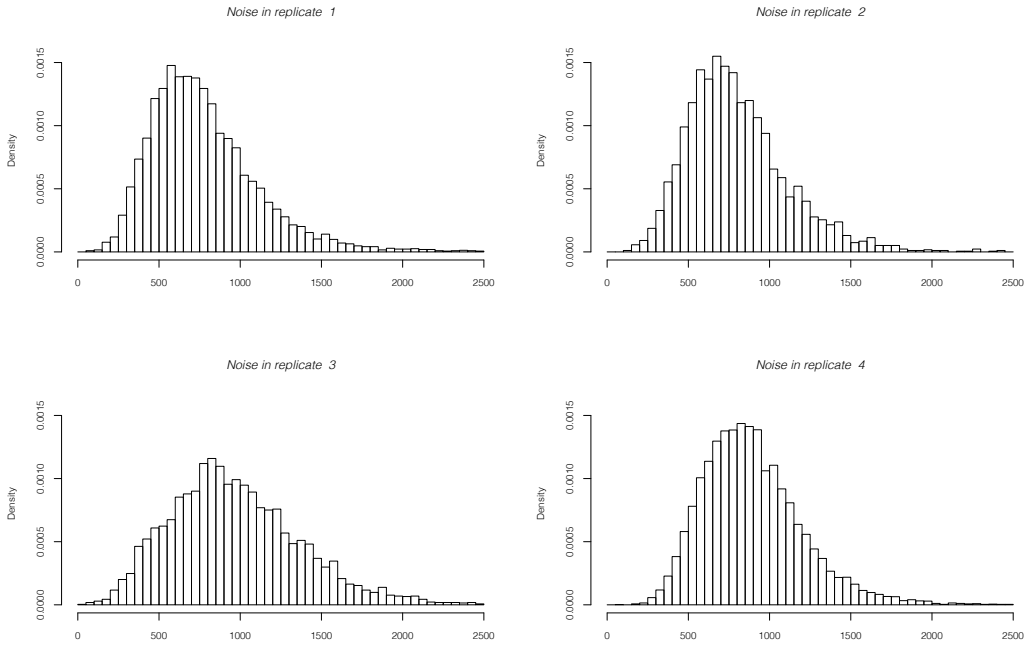


**Figure 8.2:** Histograms of the background data, taken as a proxy for the measurement error in the 4 replicates.

prior for each replicate, common for all experimental conditions measured in the same experiment. In other words, the prior obtained on the $k$-th replicate for the background data is matched with the $k$-th replicates for the experimental data, at the all levels of stimulation, for $k = 1, ..., 4$.

In particular, we use the *CRISPR/Cas9* technology of the Sigma-Aldrich® corporation to enzymatically cut out the HIV-1 env gene from the DNA of the observed cells. For the rest, the measurement collection process is repeated identically to all the other conditions, as described in Section 6.2. In this way, the original mRNA levels are approximately zero, hence observations correspond to the background noise only, i.e. to the measurement error. The number of observations available for the current analysis, for each of the four replicates, are displayed in Table 8.1.

It is reasonable to assume that the distribution of the observations for the background of the $k$-th replicate, which we define as $Z^{(k)} = (Z_1^{(k)}, ..., Z_{N_k}^{(k)})^T$, is the same as for $\epsilon^{(k)}$. Therefore, estimated posterior densities for the parameters of $Z^{(k)}$ will be used as informative priors for the measurement error parameters. Information about the four replicates is embedded in $\mathbf{Z} = (Z^{(1)}, ..., Z^{(4)})$. Figure 8.2 shows the

histograms of the observations, for the four replicates. We notice that, in some cases, the histograms are slightly asymmetric; however this phenomenon is not too evident and it may be due to an imperfect deletion of the gene of interest from the DNA. Nevertheless, assuming normally distributed errors appears to be a reasonable approximation.

We implement a simple Bayesian hierarchical model to infer parameters $\mu_\epsilon^{(k)}$ and $\sigma_\epsilon^{(k)}$, $k = 1, ..., 4$. The conjugate prior distribution choices are the same as in the general analysis, specified in Section 7.4. In particular, *a priori* $\mu_\epsilon^{(k)} \sim log\mathcal{N}(\mu_a, \frac{1}{\tau_a})$ and $\sigma_\epsilon^{(k)} \sim log\mathcal{N}(\mu_b, \frac{1}{\tau_b})$, $k = 1, ..., 4$, where $\mu_a|\tau_a, \sim \mathcal{N}(0, \frac{10^4}{\tau_a})$, $\mu_b|\tau_b, \sim \mathcal{N}(0, \frac{10^4}{\tau_b})$ and $\tau_a, \tau_b \sim \mathcal{G}(0.001, 0.001)$.

We define the hyperparameter vector for the analysis of this Section as $\Theta_\epsilon = (\Theta_{\epsilon_a}, \Theta_{\epsilon_b})$, where $\Theta_{\epsilon_a} = (\mu_a, \tau_a)^T$ and $\Theta_{\epsilon_b} = (\mu_b, \tau_b)^T$. We also gather together the hierarchical parameters in $\mu_\epsilon = (\mu_\epsilon^{(1)}, ..., \mu_\epsilon^{(4)})^T$ and $\sigma_\epsilon = (\sigma_\epsilon^{(1)}, ..., \sigma_\epsilon^{(4)})^T$.

Figure 8.3 represents the graphical model associated to the hierarchical system for the measurement error exploratory analysis.

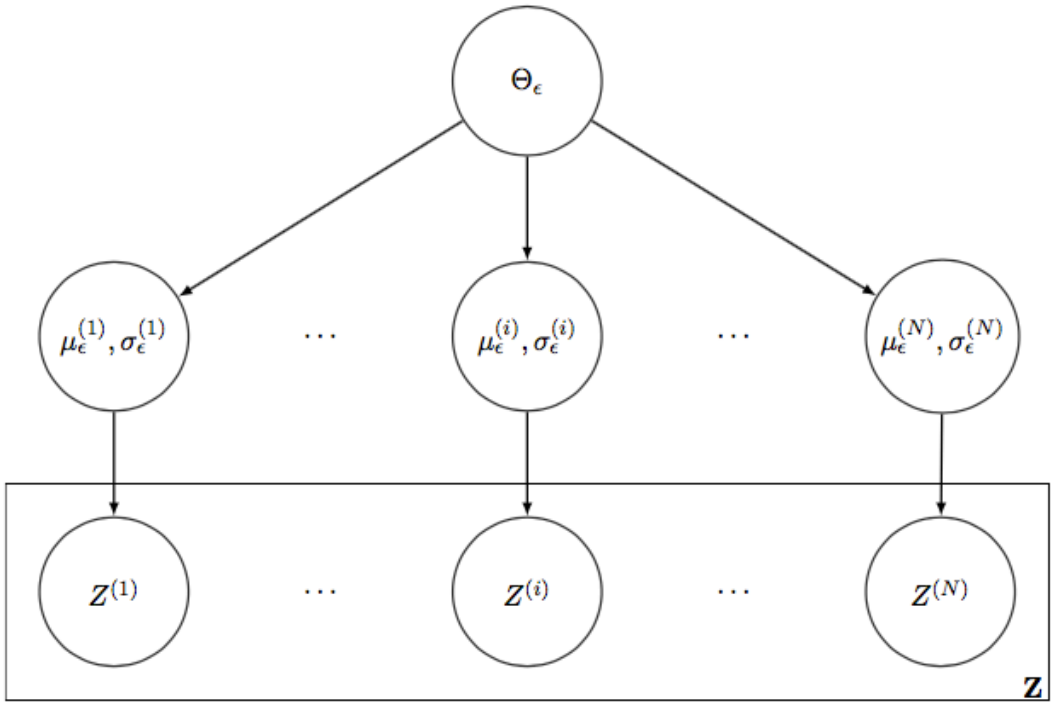


**Figure 8.3:** Graphical model for the hierarchical measurement error analysis. On the top of the graph we have the hyperparameters $\Theta_\epsilon$, which generate the hierarchical parameters: from these, the observations $Z^{(1)}, \ldots, Z^{(N)}$ are sampled.

We now present the MCMC algorithm we use to sample from the posterior densities of parameters.

**Initialisation**: First we initialise all parameters $\mu_\epsilon$, $\sigma_\epsilon$ and $\Theta_\epsilon$.

**Update**: At each iteration of the MCMC, $r = 1, ..., R$, where $R$ is the total number of iterations, we alternately sample from the conditional distributions of the hyperparameters and of the hierarchical ones, as described below.

hyperparameters: Owing to the conjugacy of the hyperprior, following the structure in Section 7.5, we use a Gibbs step to sample from the posterior distribution of the hyperparameters, conditional on the hierarchical ones:

$$\begin{aligned}
\mu_a|\tau_a &\sim \mathcal{N}\left(\frac{\tau_a \sum_{k=1}^{4} log(\mu_\epsilon^{(k)})}{\tau_a/10^4 + 4\,\tau_a}, (\tau_a/10^4 + 4\,\tau_a)^{-1}\right), \\
\tau_a|\mu_a &\sim Gamma\left(0.001 + 4/2,\, 0.001 + 1/2 \sum_{k=1}^{4} \left(log(\mu_\epsilon^{(k)}) - \mu_a\right)^2\right), \\
\mu_b|\tau_b &\sim \mathcal{N}\left(\frac{\tau_b \sum_{k=1}^{4} log(\sigma_\epsilon^{(k)})}{\tau_b/10^4 + 4\,\tau_b}, (\tau_b/10^4 + 4\,\tau_b)^{-1}\right), \\
\tau_b|\mu_b &\sim Gamma\left(0.001 + 4/2,\, 0.001 + 1/2 \sum_{k=1}^{4} \left(log(\sigma_\epsilon^{(k)}) - \mu_b\right)^2\right).
\end{aligned}$$

Hierarchical parameters: For each $k = 1, ..., 4$, we jointly sample $\mu_\epsilon^{(k)}$ and $\sigma_\epsilon^{(k)}$, via a Metropolis step; candidates are proposed from a normal distribution with adaptive covariance matrix. We call $\mu_{*\,\epsilon}^{(k)}$ and $\sigma_{*\,\epsilon}^{(k)}$ the new candidates, while $\mu_\epsilon^{(k)}$ and $\sigma_\epsilon^{(k)}$ represent the current iteration values.
The acceptance rate for $(\mu_{*\,\epsilon}^{(k)}, \sigma_{*\,\epsilon}^{(k)})$ is $min(1, A_\sigma)$, where

$$A_\sigma = \frac{\Phi(log(\mu_{*\,\epsilon}^{(k)})|\mu_a, \frac{1}{\tau_a})\, \Phi(log(\sigma_{*\,\epsilon}^{(k)})|\mu_b, \frac{1}{\tau_b}) \prod_{i=1}^{N_k} \Phi(Z_i^{(k)}|\mu_{*\,\epsilon}^{(k)}, \sigma_{*\,\epsilon}^{2\,(k)})}{\Phi(log(\mu_\epsilon^{(k)})|\mu_a, \frac{1}{\tau_a})\, \Phi(log(\sigma_\epsilon^{(k)})|\mu_b, \frac{1}{\tau_b}) \prod_{i=1}^{N_k} \Phi(Z_i^{(k)}|\mu_\epsilon^{(k)}, \sigma_\epsilon^{2\,(k)})}.$$

At both numerator and denominator, the first two elements refer to the prior densities of the logarithm of the mean and standard deviation of the measurement error. The third element represents the density for the measurement error itself, as defined in (7.18) in a non-hierarchical setting.

We apply this algorithm to our background data to infer the measurement error mean and standard deviation parameters in each replicate. Chains are run for

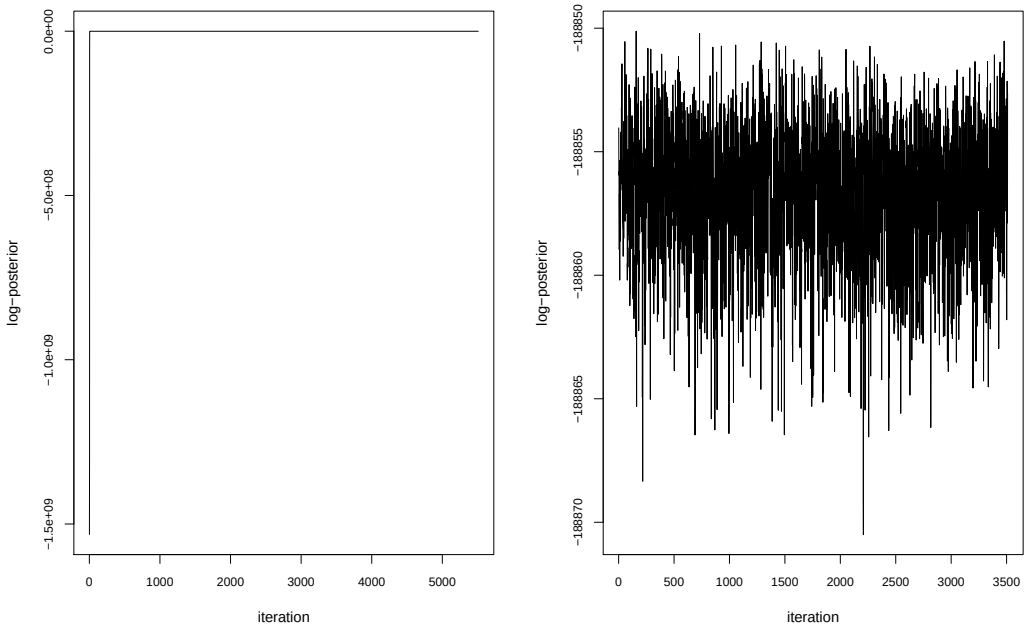


**Figure 8.4:** Trace plots for the posterior chain of the log-posterior, $log(p(\Theta_\epsilon|\mathbf{Z}))$, with $p(\Theta_\epsilon|\mathbf{Z})$ defined in (8.1), for the exploratory analysis of the measurement error. Entire chain (left plot) and chain without *burn-in* (right image).

$3.5 \times 10^5$ iterations after convergence is achieved, while the initial $2 \times 10^5$ iterations are discarded as *burn-in*. Figure 8.4 shows the trace plot for the evolution of the posterior chain, before and after convergence is reached, for the full log-posterior density, $log(p(\Theta_\epsilon|Z))$, where $p(\Theta_\epsilon|Z)$ is defined as

$$p(\Theta_\epsilon|Z) \;=\; L_Z(\mu_\epsilon, \sigma_\epsilon; \mathbf{Z})\, p(\mu_\epsilon, \sigma_\epsilon|\Theta_\epsilon)\, p(\Theta_\epsilon), \tag{8.1}$$

with

$$L_Z(\mu_\epsilon, \sigma_\epsilon; \mathbf{Z}) \;=\; \prod_{k=1}^{4} \prod_{i=1}^{N_k} \Phi(Z_i^{(k)}|\mu_\epsilon^{(k)}, \sigma_\epsilon^{2\,(k)}), \tag{8.2}$$

$$p(\mu_\epsilon, \sigma_\epsilon|\Theta_\epsilon) \;=\; \prod_{k=1}^{4} \Phi(log(\mu_\epsilon^{(k)}|\mu_a, \frac{1}{\tau_a})\, \Phi(log(\sigma_\epsilon^{(k)}|\mu_b, \frac{1}{\tau_b}) \tag{8.3}$$

$$\text{and} \qquad p(\Theta_\epsilon) \;=\; p(\Theta_{\epsilon_a})\, p(\Theta_{\epsilon_b}) \tag{8.4}$$

denoting the full likelihood, prior and hyperprior densities of the background data, respectively.

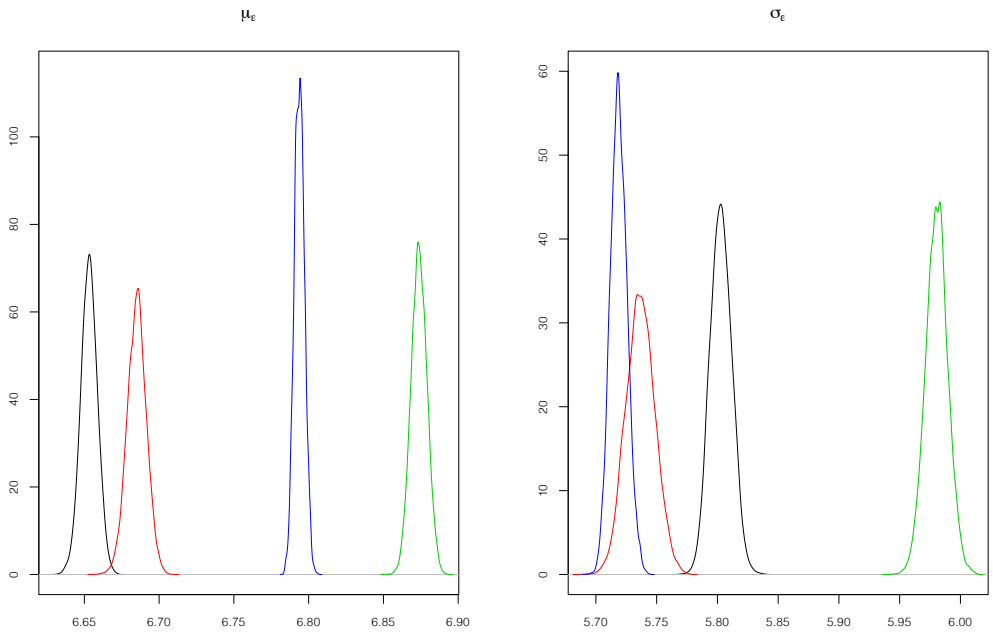


**Figure 8.5:** Posterior densities, in the log space, for the mean and standard deviation of the measurement error in the four replicates, $log(\mu_\epsilon^{(k)})$ (left plot) and $log(\sigma_\epsilon^{(k)})$ (right image), $k = 1, ..., 4$. Parameters are inferred, via a Bayesian hierarchical analysis, from the background data.

Figure B.79 in the Appendix shows the trace plots, in the log space, for the chains of the hierarchical parameters $log(\mu_\epsilon^{(k)})$ and $log(\sigma_\epsilon^{(k)})$ for all 4 replicates, $k = 1, ..., 4$, while Figure 8.5 shows the resulting posterior densities. From each one of these densities, we compute the posterior mean and standard deviation of parameters, in the log space. Values are listed in Section 7.4 in formulae (7.34)-(7.37), in particular $\mu_{\mu_\epsilon}^{(k)}$ and $\sigma_{\mu_\epsilon}^{(k)}$ indicate the posterior mean and standard deviation of $log(\mu_\epsilon^{(k)})$, respectively, while $\mu_{\sigma_\epsilon}^{(k)}$ and $\sigma_{\sigma_\epsilon}^{(k)}$ represent the posterior mean and standard deviation of $log(\sigma_\epsilon^{(k)})$, respectively, for $k = 1, ..., 4$. As explained before, these elements will be used in the next Section as a constant prior for the measurement error parameters, with a distinct prior for each replicate, where all experimental conditions for the same replicate are assumed to have, *a priori*, the same measurement error distribution.

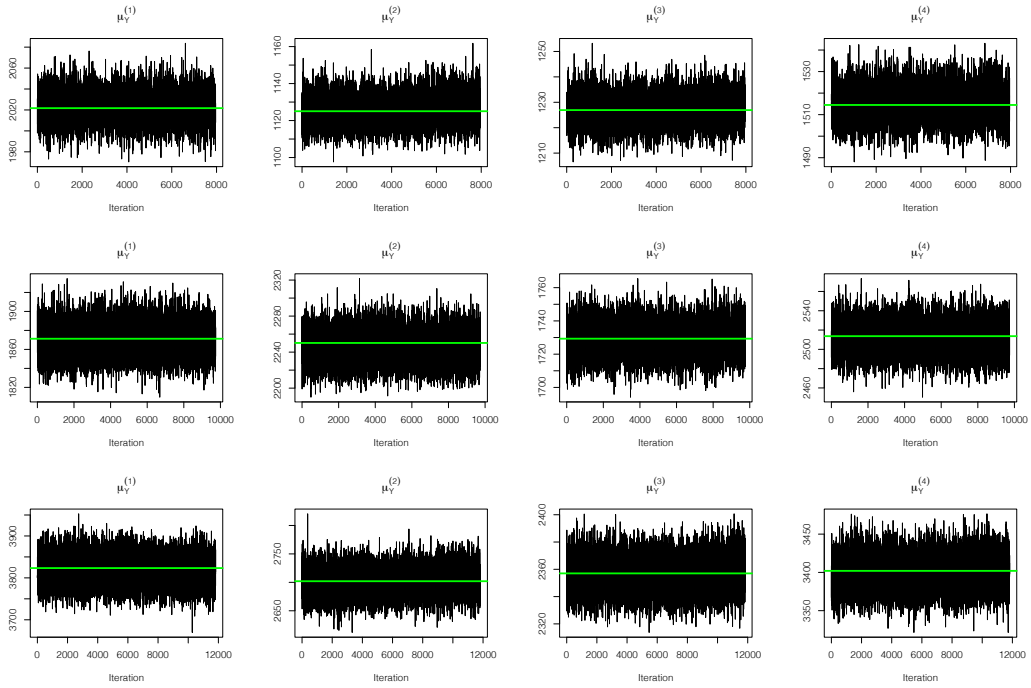


**Figure 8.6:** Trace plots for the hierarchical posterior chains of $\mu_Y^{(k)}$, for all experiments. Rows, from top to bottom, correspond to the analyses for cells under the induction by *tetracycline* at 0, 5 and 10 ng/ml, respectively. Columns, from left to right, refer to experiments 1 to 4, respectively. The horizontal green lines denote the mean of the observations in the respective experiment. Chains are thinned with a tinning factor of 100.

## 8.2 Data analysis

We apply the methodology described in Section 7.5 to our experimental data, independently for each experimental condition.

After a *burn-in* period, which is discarded, to allow for the MCMC posterior chains to reach convergence, that is assessed by eye, each analysis is run for at least $8 \times 10^5$ more iterations. The acceptance rates for blocks $b_1$, $b_2$ and $b_3$, as described in Section 7.4, are: 0.9, 9.3 and 49.8%, for the condition with no stimulation, 0.3, 18.4 and 51.7%, under the induction by 5 ng/ml of *tetracycline*, and 0.1, 8.4 and 51.6% when tetracycline level is 10 ng/ml. The low acceptance rate for the first block, resulting in highly autocorrelated chains, is increased by using a thinning factor of 100 on the posterior chains, which leads to an acceptance rate for $b_1$ of 57.6, 21.6 and 10.7% for stimulation levels at 0, 5 and 10 ng/ml, respectively. The acceptance rates for each elements of the latent states, i.e. $P_i^{(k)}$ and $X_i^{(k)}$, is approximately 31 and 89%, for the non stimulated condition, 10 and 90%, when *tetracycline* is used

| Exponential of | 0 ng/ml | | 5 ng/ml | | 10 ng/ml | |
|---|---|---|---|---|---|---|
| hypermean for | LB | UB | LB | UB | LB | UB |
| $\tilde{\alpha}_0^{(k)}$ | 1.81 | 25.76 | 0.00 | 24.95 | 0.00 | 29.52 |
| $\tilde{\alpha}_1^{(k)}$ | 2004.65 | 6355.40 | 3417.79 | 4645.52 | 1386.46 | 3801.19 |
| $\tilde{k}_{ON}^{(k)}$ | 0.04 | 0.56 | 0.22 | 1.09 | 0.59 | 1.29 |
| $\tilde{k}_{OFF}^{(k)}$ | 7.83 | 51.03 | 7.83 | 17.89 | 0.67 | 18.37 |
| $\kappa^{(k)}$ | 0.46 | 24.07 | 4.42 | 7.86 | 3.10 | 13.58 |
| $\mu_\epsilon^{(k)}$ | 722.40 | 996.60 | 721.72 | 1012.39 | 711.18 | 1005.39 |
| $\sigma_\epsilon^{(k)}$ | 278.83 | 376.12 | 267.50 | 378.93 | 272.99 | 393.91 |

**Table 8.2:** 0.95 level HPD CIs in the exponential of the hypermean parameters, for the three experimental conditions. "LB" and "UB" indicate the limits of the HPD CIs and stand for lower bound and upper bound respectively.

at 5 ng/ml, and 7 and 84%, for the highest level of induction, 10 ng/ml.

Figures B.80-B.82, in the Appendix, show the trace plots of the converged chains for the hypermean parameters in the three experimental conditions. Figure 8.6 shows, for every experimental condition, the trace plots of the converged chains for the mean of observations in each experiment, $\mu_Y^{(k)}$ defined in (7.38), for $k = 1, ..., 4$. The green horizontal line indicates the value observed in the respective sample. Figures B.83-B.85 in the Appendix show the converged trace plots of each replicate, under all experimental conditions, for both mean and standard deviation of the observed data in each experiment, $\mu_Y^{(k)}$ and $\sigma_Y^{(k)}$, $k = 1, ..., 4$, and compares them with the values observed in the corresponding samples, denoted by the green horizontal line. The second quantity is defined as $\sigma_Y^{(k)} = \sqrt{Var(Y_i^{(k)})}$ where, from the measurement equation (7.20), we have that $Var(Y_i^{(k)}) = \kappa^2 \, Var(X_i^{(k)}) + \sigma_\epsilon^2$, with $Var(X_i^{(k)})$ defined in (7.42).

We also display the posterior densities for the hypermean and precision parameters of all experimental conditions in Figures 8.7 and B.86, respectively, whilst Figure B.87 shows the posterior densities for the hierarchical parameters. Table 8.2 reports the 0.95 level highest posterior density (HPD) credible intervals (CIs) for the exponential of the hypermean parameters for all three experiments which, in log-normal distributions, correspond to the posterior modes of the hierarchical parameters. In Tables A.9-A.11 in the Appendix we also display the 95% CIs for the hierarchical parameters and for some interesting reparametrizations of such parameters which will later be described.

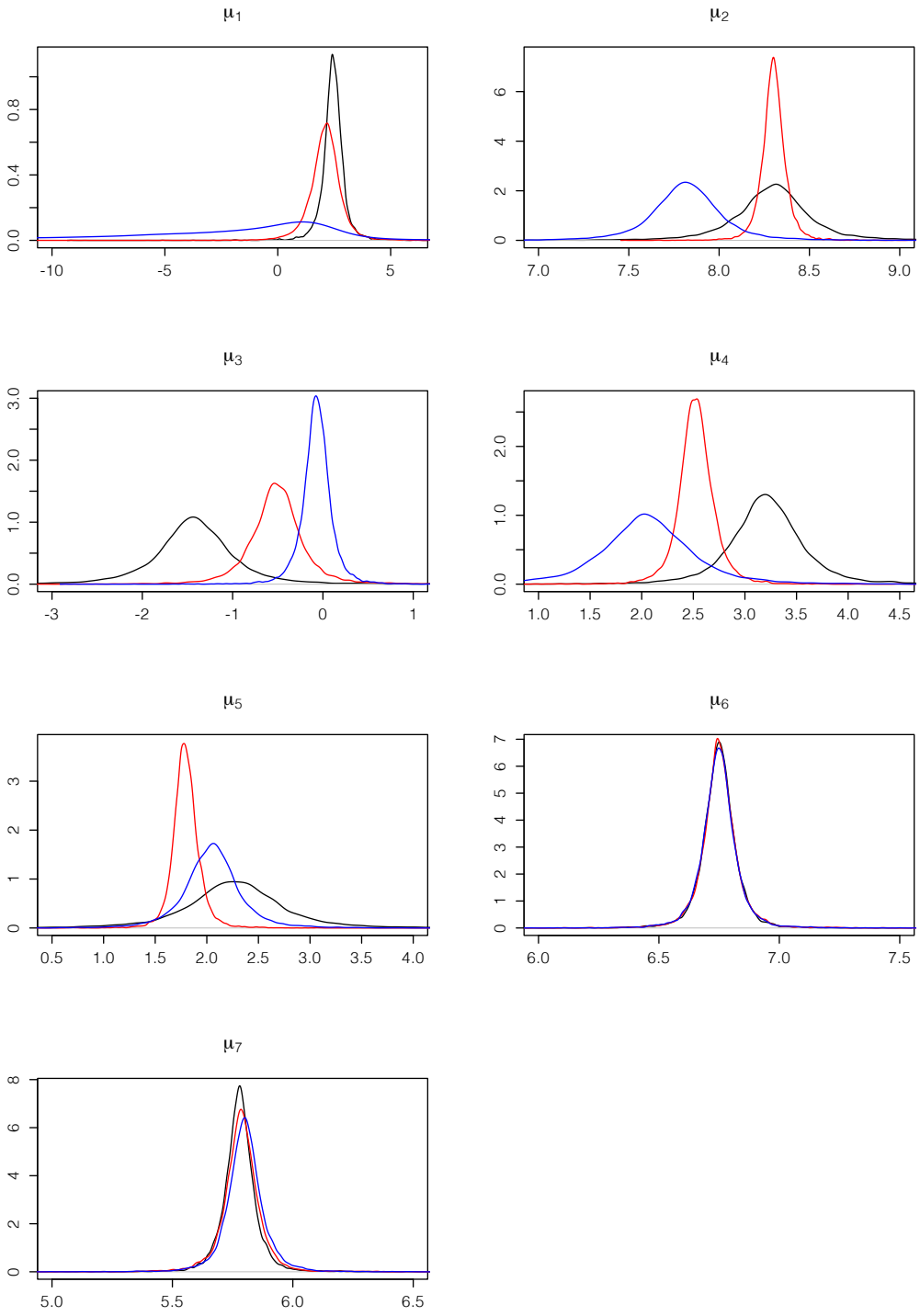


**Figure 8.7:** Posterior densities for the hypermean parameters in the three experimental conditions: the black, red and blue lines refer to data induced at 0, 5 and 10 ng/ml, respectively. Images, from left to right and top to bottom, refer in the order to the hypermeans of $\tilde{\alpha}_0^{(k)}, \tilde{\alpha}_1^{(k)}, \tilde{k}_{ON}^{(k)}, \tilde{k}_{OFF}^{(k)}, \kappa^{(k)}, \mu_\epsilon^{(k)}$ and $\sigma_\epsilon^{(k)}$.

The posterior densities for the measurement error parameters are very similar between experimental conditions, due to the informative priors used. Also the proportionality constant, $\kappa^{(k)}$, seems to be only marginally affected by the induction of cells by the stimulant. The major differences between conditions are observed in the parameters involved in the biological process, which seems reasonable as the stimulation, *a priori*, is not thought to strongly influence the measurement process. In particular, the transcription rates, $\tilde{\alpha}_0^{(k)}$ and $\tilde{\alpha}_1^{(k)}$, do not increase with stimulation; indeed, the highest induction level seems to be associated with slightly lower transcription levels. The increased levels for the observations under higher stimulation are mainly explained by the switch parameters: as stimulation increases the ON switch rate, $\tilde{k}_{ON}^{(k)}$, gradually increases, while the opposite happens to the OFF switch rate, $\tilde{k}_{OFF}^{(k)}$. Indeed, from Table 8.2 we notice that these two parameters are the only ones with non-overlapping 0.95 level HPD CIs when comparing induction levels 0 and 10 ng/ml. Therefore, it seems that the more a cell is stimulated, the more time it spends in the active state, hence generating higher levels of mRNA.

In order to further investigate these behaviours, we study some interesting reparametrizations of the hierarchical parameters. In particular, for each experiment $k = 1, ..., 4$, we consider the following quantities:

- $\frac{\tilde{\alpha}_0^{(k)}}{\tilde{\alpha}_1^{(k)}} = \frac{\alpha_0^{(k)}}{\alpha_1^{(k)}}$, i.e. the ratio between the transcription in the OFF and ON states, taken as a measure of how much more active the gene is when in the ON state as compared to the OFF state. The equivalence above is due to the fact that the degradation rate, $\beta$, is present in both $\tilde{\alpha}_0^{(k)}$ and $\tilde{\alpha}_1^{(k)}$ and hence cancels out, returning the ratio between the original transcription rates.

- $\mu_P^{(k)} = \frac{\tilde{k}_{ON}^{(k)}}{\tilde{k}_{ON}^{(k)} + \tilde{k}_{OFF}^{(k)}}$, representing the average time the gene spends in the active state.

- $\frac{1}{\tilde{k}_{ON}^{(k)}}$, indicating the mean time the gene spends in the OFF state before turning ON, where time is expressed in terms of per molecule degradation rate. Indeed such a waiting time is exponentially distributed with rate $\tilde{k}_{ON}^{(k)}$, hence its mean simply is the inverse of this parameter.

- $\frac{1}{\tilde{k}_{OFF}^{(k)}}$, representing the mean time the gene spends in the ON state before switching OFF; as above, time is expressed in terms of per molecule degradation rates. For this reason it is hard to interpret these two quantities in

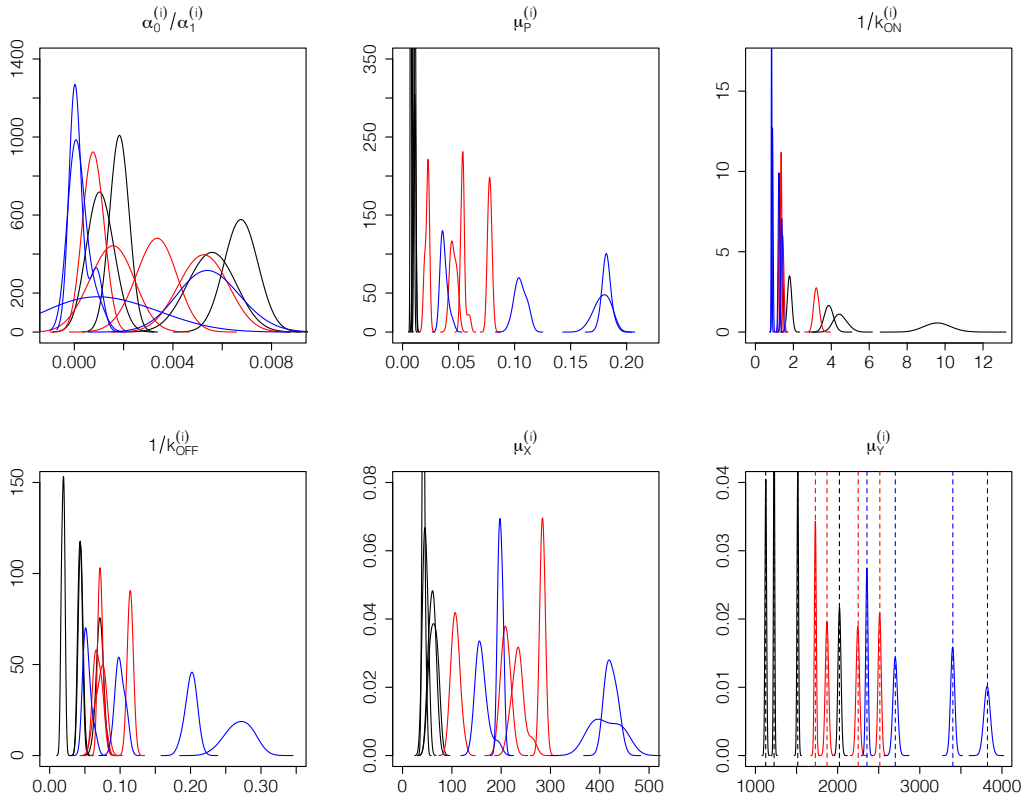


**Figure 8.8:** Posterior densities for interesting reparametrizations of the hierarchical parameters in the four experiments, under all experimental conditions: the black, red and blue lines indicate the 0, 5 and 10 ng/ml levels of induction, respectively. Each colour has four curves referring to the four experiments. In particular, from left to right and top to bottom, parameters refer to $\frac{\alpha_0^{(k)}}{\alpha_1^{(k)}}$, $\mu_P^{(k)}$, $\frac{1}{\tilde{k}_{ON}^{(k)}}$, $\frac{1}{\tilde{k}_{OFF}^{(k)}}$, $\mu_X^{(k)}$ and $\mu_Y^{(k)}$.

absolute terms; nonetheless comparisons between experimental conditions can lead to further insight into how stimulation affects the switch rates.

- $\mu_X^{(k)}$, i.e. the mean of the mRNA population at equilibrium, defined in (7.41).
- $\mu_Y^{(k)} = \kappa^{(k)}\,\mu_X^{(k)} + \mu_\epsilon^{(k)}$, representing the mean of the observed data.

Figure 8.8 displays the posterior densities for these quantities in each experiment, under all experimental conditions. We also report the respective 0.95 level HPD CIs in each replicate in Tables A.9-A.11 in the Appendix. In order to better visualize the differences between conditions, we average each of these quantities across the four replicates and plot the respective posterior densities in Figure 8.9. From

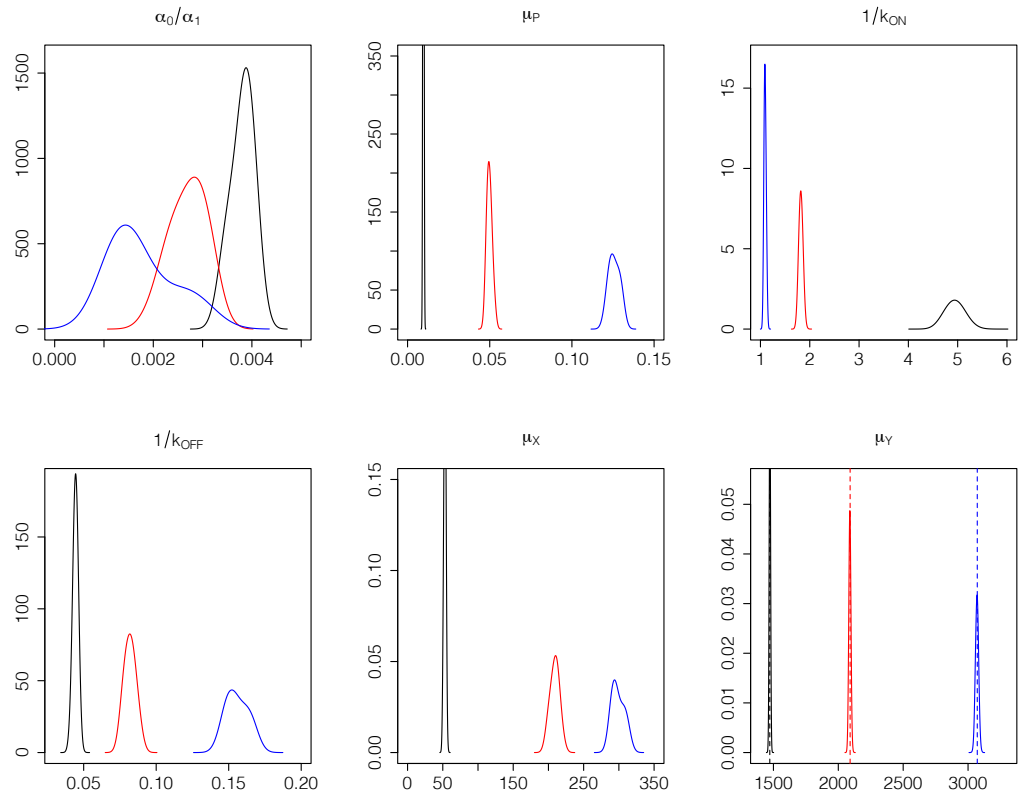


**Figure 8.9:** The images refer to the same elements as in Figure 8.8; now for each experimental condition, the posterior densities refer to the average across the four experiments.

the analyses of these figures we notice that the ON state is much more active than the OFF state: $\alpha_0$ is only 0-0.5% the value of $\alpha_1$; furthermore the results suggest that the value of $\alpha_0/\alpha_1$ might decrease for higher levels of stimulation. The most evident distinction between conditions concerns the average time the gene spends in the active state: we observe a clear increase as the level of induction goes from 0 to 5 and from 5 to 10 ng/ml. Moreover, we also find that this distinction is due to a variation in both switch rates: as the induction of *tetracycline* increases the gene spends, on average, more time in the active state before turning OFF and, conversely, spends, on average, less time in the OFF state before switching ON again. Therefore higher levels of stimulation are associated to both more frequent and longer bursts. As a result, the mean population of mRNA molecules at equilibrium sharply augments with the stimulation level. This directly affects in the average level of observations, $\mu_Y^{(k)}, k = 1, ..., 4$, which closely matches the respective value observed in the experimental data, i.e. $\left\langle Y^{(k)} \right\rangle, k = 1, ..., 4$.

Furthermore, it is interesting to notice that, in all cases, the gene spends most of the time, between 80 and 100%, in the inactive state; this is particularly true for the basal, non-stimulated, condition. This finding confirms the initial theory that transcription follows short and intense bursts. Moreover, as evident from Figures 8.8 and 8.9, such bursts also appear to be more frequent and to last for longer at higher levels of stimulation. Another interesting result concerns the average value of the mRNA population at equilibrium, which goes from a few dozens under the basal condition, to a few hundreds, when stimulated.

In the next Section we will study the fit of the model by analysing the behaviour of some residuals obtained from the difference between the observations and the latent states for the mRNA population, rescaled by the proportionality constant.

## 8.3 Residual analysis

Similar to the residual analysis performed in Part I, we study the difference between the observations and the latent data for the mRNA population, multiplied by the respective proportionality constant. In particular, in each experimental condition, we define the residuals for the $k$-th experiment, $i$-th observation and the $j$-th MCMC iteration as

$$r_i^{(k)^j} = Y_i^{(k)} - \kappa^{(k)^j} X_i^{(k)^j} - \mu_\epsilon^{(k)^j},$$

with $k = 1, ..., 4$, $i = 1, ..., N_k$ and $j = 1, ..., R$.

We also compute the standardized residuals dividing by the standard deviation of the measurement error at the respective iteration of the MCMC, $\tilde{r}_i^{(k)^j} = \frac{r_i^{(k)^j}}{\sigma_\epsilon^{(k)^j}}$, and compare these quantities with the standard normal distribution. In particular, we group the standardized residuals for $k$-th experiment in the $J \times N_k$ matrix $\tilde{r}^{(k)} = (\tilde{r}_1^{(k)}, ..., \tilde{r}_{N_k}^{(k)})$, where $\tilde{r}_i^{(k)} = (\tilde{r}_i^{(k)^1}, ..., \tilde{r}_i^{(k)^J})^T$, $i = 1, ..., N_k$. In Figure 8.10, we plot the histograms of $\tilde{r}^{(k)}$, for each experiment $k = 1, ..., 4$, obtained in the three analyses on each experimental condition. We also add a curve for the density of the standard normal distribution for visual comparison. All histograms are centred in 0, symmetric and very close to the standard normal density line. We only observe a minor deviance from normality in the third experiment of the non-stimulated condition in the first row. On the overall, the behaviour of residuals is highly satisfactory and modelling assumptions appear to be respected.

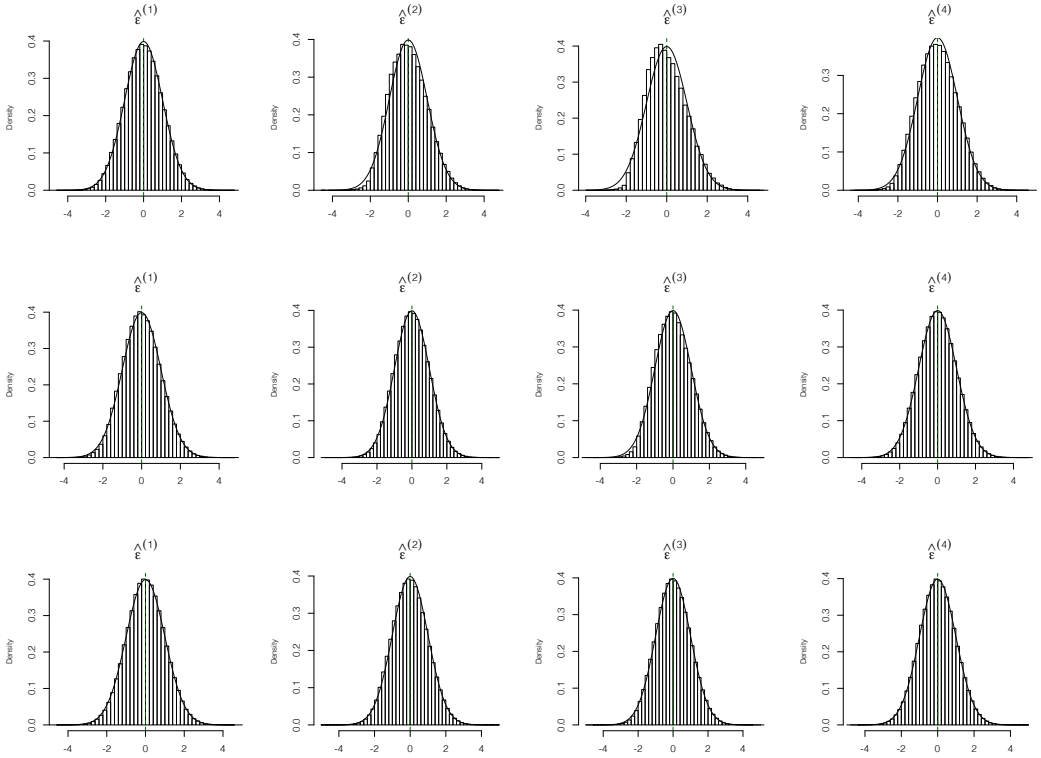


**Figure 8.10:** Histograms for the residuals in all each experiment, $\tilde{r}^{(k)}$, for all experimental conditions. The top row refers to the non-stimulated condition, the second row to the induction by *tetracycline* at 5 ng/ml and the bottom row to the induction at 10 ng/ml. The four columns refer, in the order, to the four experiments. The black line on top of the histograms indicates the density of a $\mathcal{N}(0, 1)$ distribution.

## 8.4 Summary

This Chapter has focused on the analysis of the available experimental data for the mRNA levels for the HIV-1 env gene. First we carried out an exploratory study on background noise data which allow us to infer the measurement error mean and standard deviation and, hence, to formulate an informative prior for these two parameters in each experiment.

We then applied the Bayesian hierarchical inferential methodology described in Section 7.5 to our experimental data to infer the posterior densities of parameters. By comparing results under different experimental conditions, we find that the higher level for the observations that the stimulant induces is mostly explained by changes in the switch rates. In particular, as the induction level increases, the gene spends significantly more time in the active state, where it transcribes mRNA at

a significantly higher rate. Furthermore, results indicate that transcription in the inactive state is much smaller than in the active one, with $\alpha_0$ being approximately 0-0.5% with respect to the value of $\alpha_1$. We also inferred, for each experiment, the mean time the gene is active and the average mRNA population at equilibrium; interestingly we find that, in all conditions, the gene is mostly inactive, confirming the theory that transctiption mostly happens in short and intense bursts.

Finally, we studied the standardized residuals, obtained from the observations and the latent states, and show that they fulfil the modelling assumptions as their histograms are very close to a standard normal distribution.

# Conclusions

# Chapter 9

# Possible extensions and conclusions

> Maybe the journey isn't so much about becoming anything. Maybe it's about unbecoming everything that isn't you so you can be who you were meant to be in the first place.
>
> *Unknown*

This thesis focused on two analyses of mRNA and protein populations in single cells. In both cases, we build a Bayesian hierarchical inferential framework to infer the parameters of the respective stochastic system. Methods were first validated in simulation studies and then applied to the available experimental data, contributing to gaining a better understanding of the biological mechanisms under study. In this Chapter we briefly summarize the achievements of the thesis and discuss alternative approaches which may have been used and possible extensions of our work.

## 9.1 Possible extensions and alternative approaches

In both analyses we use a measurement equation to relate the observations to the original unobserved molecular populations, involving a proportionality constant and a Gaussian error. Alternative assumptions might also be investigated to improve model realism. For instance, for the analysis in Part II, one may also employ a strictly positive distribution for the error such as the Gamma or the log-normal.

Furthermore, in order to deal with the latent states, we use data augmentation procedures to sample the unobserved data. Alternatively, one could use particle approaches, e.g. particle marginal Metropolis-Hastings (PMMH) where the marginal density of observations is replaced by a Monte Carlo (MC) estimate. In Part I, we approximate the exact likelihood of the process with the diffusion and Euler-Maruyama approximations; alternatively, one could proceed via the linear noise approximation (LNA) as, for instance, in Finkenstädt *et al.* (2013).

Further work will be carried out by applying the methodology we proposed in Part II to other experimental conditions; more gene expression data from the HIV-1 *env* gene will be analysed, i.e. with and without a point mutation of a $pA+$ site and under various levels of induction by *tetracycline*. This point mutation is believed to disrupt gene loops (Mapendano *et al.*, 2010) and is typically associated with lower mRNA transcription. Gene loops are three dimensional interactions of the DNA which bring start and end sites of genes closer to each other and, indirectly, facilitate transcription by allowing for polymerase recycling (Hampsey *et al.*, 2011; Schleif, 1992; Hebenstreit, 2013). Comparing inference in the presence and absence of the mutation will allow us to obtain indirect information about the mechanism of gene loops, assuming the absence of the mutation to be a proxy for the presence of gene loops. Two further characteristics of the cells will also be available: the cell size and the phase of the cell cycle. By applying our inferential framework to expression levels stratified with respect to these two covariates, we would be able to observe how parameters vary during the cell cycle and for cells of different sizes. Moreover, the same type of gene expression data will also be available for a second gene, namely hemoglobin subunit beta (HBB), also known as beta globin. We will also analyse these data to compare inference between experimental conditions on a second gene.

## 9.2 Summary of achievements and conclusions

The work presented in this thesis has been motivated by two specific biological systems; therefore, the methodologies described have been developed with a focus on the available experimental data. Nevertheless, the methods and algorithms proposed can be applied to a wide variety of contexts. Here we summarize the main achievements of this thesis, separating them in two groups.

Concerning the methodological side, we propose two inferential methods and corresponding algorithms, which can be used to retrieve the parameters of interest in problems, in both longitudinal and stationary contexts, where the data is characterized by a source of random noise. Although such methods have been tuned

specifically for the available data, they can be easily generalised and employed in a broad class of problems. Furthermore, for the two-state switch model of Part II, we prove that the mRNA stationary distribution of our model corresponds to a Poisson-beta. This finding can improve model realism and ease inference on gene expression from switch models. Indeed it facilitates the use of the two-state switch model with transcription in both states, which represents a more realistic extension of the simpler model with transcription in the ON state only.

A second contribution of this thesis concerns the biological field. Firstly, we have developed improved systems and reaction networks for the protein and mRNA populations under study. Secondly, by means of the mechanistic modelling approaches, combined with exploratory studies and with the full Bayesian hierarchical analyses, we obtained a deeper understanding of the mechanisms of the two systems under study and deduced how molecular processes vary when cells are induced by a stimulant.

We find that the Nrf2 system behaves, in most cases, as a noise-induced oscillator. Therefore, the oscillations we observe in the experimental data ought to be attributed to the stochasticity induced by the limited molecular populations in single cells and would fade out in a deterministic context. We also find that, as expected, this noise-induced oscillatory behaviour is more evident in stimulated cells and that a key component of this mechanism is represented by the delay term. Furthermore, we show that the nuclear export of Nrf2 is significantly faster than its nuclear import. We also obtained information about interesting quantities, such as the ratio between the cytoplasmic and the nuclear volumes in a cell. Moreover parameters' posterior densities show interesting differences between experimental conditions. In particular, induced cells display faster import and export rates as compared to cells under the basal condition; this finding is compatible with the more dynamic and oscillatory pattern exibited by the experimental data when stimulated.

Our major biological finding in the data analysis in Part II is that the higher gene expression, observed in cells induced by *tetracycline*, is mostly explained by changes in the switch rates. In particular, the higher the level of stimulation, which translates into higher mRNA populations, the longer the gene spends in the active state, whilst transcription rates themselves do not appear to increase with the level of *tetracycline*. We also find that genes are mostly inactive, spending only a small fraction of time in the active state, about 1-20% depending on the level of stimulation. Furthermore, transcription in the OFF state is inferred to be much lower than in the ON state; in particular, the ratio between the former and the latter is approximately 0-0.5%. These findings confirm the prior hypothesis that transcription follows short

and intense bursts. The analysis of further available data, on different experimental conditions and on a new gene, will contribute to increase the understanding of this transcriptional mechanism and the role of gene loops.

A final remark concerns the use of a stochastic approach as opposed to deterministic modelling: our analyses highlighted the importance of modelling single cell processes in a stochastic framework. Indeed biological processes in individual cells, such as transcription or translation, are inherently stochastic (Singh *et al.*, 2013) and ought to be treated as such. For instance, we showed that the Nrf2 system can only oscillate when induced by noise. In Part II we also saw how the mRNA populations can only be appropriately represented by a model which accounts for the wide and over-dispersed structure of the data, whose mean alone would be a highly simplistic summary. Moreover, the use of a stochastic approach has allowed us to infer more parameters than we would have been able to in a deterministic framework. Indeed, in both analyses, we were able to estimate the measurement equation proportionality constants as well as the measurement error variances and, hence, to separate the effect of these nuissance parameters from that of the biological quantities of interest. Finally, quantifying the variance of these systems, which would be impossible in a deterministic scenario, is extremely important when studying the effect of interventions on the system, such as the introduction of a drug or the repression of a protein.

## 9.3 Software

The software used for all analyses shown in the thesis has been  (R Core Team, 2016). Although the vast majority of the code is self written, the following packages have been used for specific computations: *deSolve* (Soetaert *et al.*, 2016), *parallel* (R Core Team, 2016), *doParallel* (Calaway *et al.*, 2016), *coda* (Plummer *et al.*, 2016), *MCMCglmm* (Hadfield, 2016) and *mvtnorm* (Genz *et al.*, 2016).

# Appendix A

# Tables

## A.1 Part I

| Cell | $k_d^{(i)}10^7$ | $k_a^{(i)}10^5$ | $K_a^{(i)}$ | $\mu_\tau^{(i)}$ | $\gamma^{(i)}10^4$ | $\delta^{(i)}10^{11}$ | $c^{(i)}$ | $\kappa_N^{(i)}10^{-4}$ | $\sigma_N^{(i)}$ | $\sigma_C^{(i)}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.5 | 31.6 | 1.9 | 27.6 | 6.0 | 16.7 | 8.7 | 1.8 | 67.9 | 16.6 |
| 2 | 429.7 | 73.2 | 0.8 | 7.0 | 3.1 | 9.1 | 14.2 | 5.3 | 15.2 | 15.4 |
| 3 | 118.5 | 6.1 | 26.0 | 11.2 | 3.1 | 12.8 | 13.2 | 1.5 | 18.2 | 26.6 |
| 4 | 27.2 | 53.7 | 1020.8 | 10.2 | 6.2 | 12.0 | 14.1 | 0.6 | 5.9 | 36.2 |
| 5 | 105.7 | 5.3 | 0.0 | 9.9 | 6.8 | 16.0 | 10.0 | 4.9 | 41.9 | 28.3 |
| 6 | 9.0 | 7.8 | 480.1 | 9.4 | 9.8 | 19.8 | 15.6 | 3.3 | 21.9 | 28.4 |
| 7 | 35.6 | 108.7 | 34.3 | 43.9 | 5.3 | 18.3 | 9.8 | 3.8 | 38.4 | 65.5 |
| 8 | 1002.4 | 288.7 | 2746.4 | 6.0 | 6.4 | 12.4 | 15.1 | 1.7 | 13.5 | 21.9 |
| 9 | 21.8 | 4.4 | 0.4 | 24.7 | 9.3 | 19.0 | 10.2 | 0.7 | 23.0 | 7.8 |
| 10 | 18.9 | 42.3 | 13.3 | 27.5 | 30.4 | 15.2 | 7.9 | 3.2 | 22.4 | 34.7 |
| 11 | 125.2 | 26.6 | 0.0 | 1.4 | 5.1 | 15.2 | 10.1 | 5.2 | 25.5 | 36.5 |
| 12 | 488.2 | 22.6 | 2.7 | 23.2 | 3.1 | 12.5 | 5.3 | 1.6 | 17.7 | 28.2 |
| 13 | 9.7 | 191.9 | 136.6 | 8.7 | 6.9 | 17.7 | 10.3 | 1.4 | 14.4 | 5.3 |
| 14 | 15.5 | 420.4 | 20.2 | 5.7 | 19.4 | 18.6 | 8.4 | 2.1 | 1.5 | 2.5 |
| 15 | 304.8 | 37.6 | 22.7 | 7.0 | 7.2 | 8.7 | 11.4 | 2.4 | 11.4 | 15.5 |
| 16 | 17.4 | 96.3 | 143.5 | 35.1 | 10.9 | 16.8 | 12.0 | 4.2 | 22.0 | 8.8 |
| 17 | 6.7 | 39.6 | 0.4 | 19.9 | 7.2 | 22.0 | 17.7 | 2.9 | 2.7 | 9.2 |
| 18 | 164.8 | 11.0 | 0.7 | 11.9 | 17.6 | 11.5 | 15.4 | 4.1 | 60.5 | 10.2 |
| 19 | 39.8 | 65.8 | 3.7 | 16.9 | 2.7 | 32.1 | 6.4 | 2.6 | 78.9 | 35.2 |
| 20 | 141.7 | 12.1 | 0.0 | 7.1 | 3.1 | 24.3 | 16.0 | 3.5 | 10.6 | 20.6 |
| 21 | 274.2 | 42.2 | 0.2 | 5.3 | 12.0 | 18.8 | 10.3 | 7.3 | 23.3 | 18.4 |
| 22 | 19.6 | 92.9 | 0.7 | 14.2 | 3.8 | 12.4 | 43.3 | 2.8 | 5.5 | 15.1 |
| 23 | 1081.5 | 78.2 | 0.2 | 7.7 | 6.7 | 14.2 | 10.2 | 2.0 | 6.5 | 6.0 |
| 24 | 38.4 | 8.7 | 18.8 | 5.0 | 10.1 | 14.5 | 14.1 | 4.6 | 11.4 | 29.3 |
| 25 | 2.6 | 1036.6 | 11.1 | 16.5 | 9.1 | 9.1 | 20.4 | 2.8 | 23.4 | 26.6 |
| 26 | 24.9 | 12.6 | 0.0 | 31.5 | 10.1 | 15.5 | 15.6 | 3.3 | 7.3 | 3.3 |
| 27 | 128.2 | 46.2 | 0.3 | 9.2 | 12.6 | 19.1 | 5.8 | 2.7 | 73.1 | 25.6 |
| 28 | 46.5 | 39.2 | 0.4 | 6.7 | 4.1 | 13.3 | 9.5 | 2.0 | 27.5 | 24.4 |
| 29 | 15.5 | 42.5 | 72.8 | 20.6 | 4.5 | 24.8 | 11.8 | 7.9 | 55.4 | 236.8 |
| 30 | 61.6 | 2.1 | 46.3 | 11.6 | 10.5 | 24.8 | 17.2 | 2.7 | 9.4 | 3.3 |
| 31 | 1366.6 | 28.0 | 2.3 | 16.9 | 6.4 | 17.2 | 6.4 | 1.6 | 6.3 | 4.8 |
| 32 | 78.0 | 40.1 | 13.0 | 4.2 | 3.8 | 29.3 | 4.1 | 3.4 | 5.7 | 75.0 |
| 33 | 150.1 | 51.4 | 14.2 | 19.8 | 9.3 | 13.6 | 26.6 | 1.7 | 73.0 | 59.2 |
| 34 | 201.9 | 11.5 | 0.2 | 12.8 | 5.3 | 14.8 | 4.8 | 4.7 | 50.5 | 90.7 |
| 35 | 835.5 | 177.1 | 6504.2 | 10.8 | 15.9 | 17.2 | 11.4 | 1.1 | 2.0 | 17.3 |

**Table A.1:** Randomly drawn hierarchical parameter values used in the first Nrf2 simulation study.

| Cell | $k_d^{(i)}10^7$ | $k_a^{(i)}10^5$ | $K_a^{(i)}$ | $\mu_\tau^{(i)}$ | $\gamma^{(i)}10^4$ | $\delta^{(i)}10^{11}$ | $c^{(i)}$ | $\kappa_N^{(i)}10^{-4}$ | $\sigma_N^{(i)}$ | $\sigma_C^{(i)}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 48.6 | 27.5 | 2.9 | 11.5 | 6.8 | 14.9 | 9.7 | 1.6 | 12.7 | 8.1 |
| 2 | 29.4 | 29.0 | 4.7 | 7.8 | 10.0 | 19.1 | 12.7 | 2.2 | 11.8 | 15.9 |
| 3 | 42.5 | 31.5 | 5.4 | 14.0 | 5.5 | 20.3 | 18.3 | 2.0 | 12.2 | 15.4 |
| 4 | 27.9 | 28.0 | 3.3 | 13.0 | 7.1 | 23.8 | 16.1 | 2.8 | 10.7 | 11.3 |
| 5 | 40.5 | 24.5 | 4.2 | 10.7 | 4.7 | 11.6 | 15.6 | 2.7 | 10.6 | 13.3 |
| 6 | 37.6 | 29.9 | 3.8 | 12.7 | 5.2 | 11.2 | 13.0 | 1.8 | 10.8 | 12.9 |
| 7 | 35.4 | 39.8 | 2.9 | 12.1 | 9.3 | 19.0 | 9.8 | 3.3 | 19.7 | 13.2 |
| 8 | 30.9 | 38.1 | 4.2 | 13.3 | 5.1 | 17.5 | 10.3 | 2.5 | 23.7 | 9.6 |
| 9 | 36.5 | 23.3 | 5.3 | 12.0 | 6.3 | 14.7 | 9.8 | 1.9 | 9.6 | 13.3 |
| 10 | 34.5 | 33.5 | 4.1 | 13.2 | 3.9 | 22.7 | 11.3 | 2.7 | 16.7 | 17.9 |
| 11 | 36.3 | 31.3 | 6.0 | 15.5 | 7.5 | 16.4 | 10.7 | 3.0 | 12.2 | 9.9 |
| 12 | 37.3 | 23.5 | 4.3 | 11.3 | 4.3 | 14.0 | 9.4 | 2.0 | 14.2 | 10.1 |
| 13 | 29.2 | 26.3 | 4.9 | 13.6 | 7.7 | 10.0 | 9.3 | 2.0 | 9.6 | 11.7 |
| 14 | 30.9 | 18.3 | 2.9 | 16.4 | 7.2 | 23.7 | 11.7 | 2.7 | 15.6 | 13.1 |
| 15 | 44.8 | 35.8 | 4.3 | 12.0 | 3.9 | 13.8 | 11.4 | 2.1 | 15.6 | 11.8 |
| 16 | 27.3 | 30.0 | 6.3 | 7.8 | 6.0 | 13.5 | 10.6 | 2.1 | 18.6 | 13.9 |
| 17 | 39.7 | 19.7 | 5.4 | 8.1 | 10.1 | 17.6 | 10.9 | 2.3 | 11.4 | 7.6 |
| 18 | 30.9 | 37.6 | 5.8 | 8.9 | 6.6 | 13.9 | 13.2 | 3.0 | 19.7 | 15.0 |
| 19 | 31.2 | 15.6 | 4.6 | 10.0 | 5.3 | 27.5 | 8.6 | 1.8 | 9.5 | 16.8 |
| 20 | 39.9 | 48.8 | 4.3 | 9.6 | 5.2 | 16.8 | 13.2 | 2.5 | 11.3 | 9.4 |
| 21 | 45.6 | 38.3 | 3.9 | 9.1 | 8.5 | 14.8 | 10.8 | 1.9 | 15.6 | 12.0 |
| 22 | 34.9 | 28.0 | 3.6 | 13.4 | 8.3 | 11.1 | 12.9 | 3.2 | 9.3 | 11.0 |
| 23 | 51.0 | 24.1 | 4.1 | 17.0 | 6.0 | 14.7 | 13.4 | 1.8 | 15.4 | 16.4 |
| 24 | 36.3 | 47.4 | 4.1 | 14.6 | 6.2 | 16.3 | 18.9 | 2.4 | 18.8 | 12.7 |
| 25 | 33.3 | 23.7 | 5.9 | 12.5 | 3.3 | 15.4 | 10.0 | 3.1 | 10.9 | 14.3 |
| 26 | 28.6 | 44.2 | 5.0 | 13.8 | 7.3 | 17.3 | 13.0 | 2.3 | 14.8 | 15.5 |
| 27 | 42.2 | 22.9 | 5.8 | 6.8 | 7.0 | 18.9 | 20.3 | 3.1 | 8.9 | 9.2 |
| 28 | 36.8 | 24.5 | 7.4 | 8.7 | 6.6 | 19.1 | 13.2 | 3.1 | 14.5 | 8.8 |
| 29 | 58.8 | 21.9 | 4.0 | 13.5 | 5.5 | 13.1 | 12.6 | 3.6 | 13.9 | 14.1 |
| 30 | 35.6 | 21.1 | 4.3 | 14.9 | 8.1 | 12.0 | 16.0 | 2.7 | 13.4 | 13.3 |
| 31 | 26.4 | 15.9 | 2.2 | 13.6 | 6.9 | 11.4 | 11.4 | 2.8 | 9.6 | 8.5 |
| 32 | 36.5 | 31.9 | 2.9 | 17.7 | 4.5 | 24.8 | 10.1 | 2.3 | 11.1 | 13.3 |
| 33 | 43.5 | 21.5 | 6.0 | 11.9 | 7.4 | 18.1 | 12.1 | 2.6 | 12.9 | 9.8 |
| 34 | 36.6 | 18.4 | 3.6 | 15.7 | 7.5 | 10.3 | 14.2 | 2.8 | 11.5 | 8.9 |
| 35 | 55.3 | 22.7 | 4.1 | 13.3 | 4.1 | 16.0 | 12.5 | 4.4 | 13.0 | 12.1 |

**Table A.2:** Randomly drawn hierarchical parameter values used in the second Nrf2 simulation study.

| Cell | $k_d^{(i)}10^7$ | $k_a^{(i)}10^5$ | $K_a^{(i)}$ | $\mu_\tau^{(i)}$ | $\gamma^{(i)}10^4$ | $\delta^{(i)}10^{11}$ | $c^{(i)}$ | $\kappa_N^{(i)}10^{-4}$ | $\sigma_N^{(i)}$ | $\sigma_C^{(i)}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 668.0 | 26.3 | 2.9 | 4.6 | 298.0 | 100.0 | 29.5 | 0.8 | 9.7 | 114.2 |
| 2 | 1087.0 | 49.0 | 4.7 | 16.3 | 843.5 | 88.5 | 18.4 | 0.4 | 5.8 | 29.7 |
| 3 | 6542.6 | 31.5 | 6.0 | 20.3 | 408.7 | 166.6 | 8.2 | 0.3 | 93.1 | 22.3 |
| 4 | 1811.4 | 15.1 | 1.6 | 36.8 | 431.9 | 133.8 | 6.9 | 2.4 | 4.1 | 129.2 |
| 5 | 1352.5 | 15.8 | 5.2 | 6.7 | 278.8 | 68.2 | 10.3 | 0.7 | 2.5 | 95.3 |
| 6 | 3219.0 | 19.4 | 3.4 | 7.2 | 468.3 | 58.0 | 11.0 | 1.6 | 20.3 | 90.0 |
| 7 | 1293.5 | 22.7 | 1.7 | 33.8 | 525.3 | 102.2 | 16.7 | 0.4 | 45.8 | 4.8 |
| 8 | 1752.7 | 37.1 | 14.7 | 10.8 | 228.1 | 122.6 | 10.3 | 0.2 | 19.7 | 4.7 |
| 9 | 1960.9 | 7.2 | 4.1 | 10.8 | 220.4 | 75.6 | 12.6 | 0.4 | 6.4 | 13.2 |
| 10 | 783.9 | 64.1 | 2.5 | 24.6 | 213.8 | 88.9 | 24.2 | 0.6 | 5.5 | 29.5 |
| 11 | 1068.7 | 61.1 | 2.0 | 12.4 | 146.1 | 92.1 | 8.6 | 0.1 | 23.9 | 18.3 |
| 12 | 5832.4 | 51.5 | 6.3 | 23.5 | 199.5 | 132.9 | 18.0 | 0.2 | 5.8 | 34.5 |
| 13 | 3190.6 | 72.2 | 4.2 | 3.0 | 490.9 | 118.4 | 20.9 | 0.3 | 6.1 | 17.5 |
| 14 | 6465.9 | 30.1 | 15.5 | 3.4 | 426.4 | 116.9 | 17.2 | 0.4 | 15.9 | 15.4 |
| 15 | 2749.1 | 10.3 | 3.2 | 6.3 | 420.4 | 107.8 | 6.6 | 0.2 | 38.2 | 11.6 |
| 16 | 4577.4 | 7.9 | 4.1 | 26.2 | 527.1 | 52.5 | 25.5 | 0.3 | 42.8 | 48.5 |
| 17 | 2047.1 | 14.3 | 2.8 | 17.2 | 354.9 | 99.5 | 8.1 | 0.0 | 12.8 | 5.2 |
| 18 | 2256.3 | 27.9 | 0.9 | 2.3 | 662.6 | 90.7 | 12.2 | 0.3 | 45.4 | 64.7 |
| 19 | 1611.3 | 27.5 | 5.1 | 12.0 | 136.7 | 229.2 | 6.2 | 0.3 | 53.5 | 31.7 |
| 20 | 1190.4 | 50.6 | 4.4 | 17.0 | 890.6 | 110.7 | 9.5 | 0.5 | 4.7 | 10.0 |
| 21 | 3146.6 | 33.9 | 6.8 | 14.4 | 316.8 | 132.3 | 4.2 | 0.5 | 21.9 | 55.8 |
| 22 | 3095.7 | 22.9 | 3.4 | 5.8 | 492.2 | 126.8 | 5.1 | 1.3 | 45.9 | 41.7 |
| 23 | 715.9 | 30.6 | 2.3 | 6.7 | 512.8 | 165.2 | 20.7 | 0.6 | 6.5 | 35.0 |
| 24 | 3954.3 | 17.7 | 5.8 | 10.5 | 289.4 | 112.8 | 11.5 | 0.0 | 10.0 | 25.2 |
| 25 | 2314.1 | 18.0 | 3.9 | 11.2 | 315.8 | 134.5 | 5.1 | 1.0 | 21.2 | 2.1 |
| 26 | 1436.8 | 34.2 | 18.6 | 6.3 | 823.2 | 70.3 | 12.0 | 0.2 | 39.3 | 66.6 |
| 27 | 988.5 | 35.8 | 9.1 | 7.5 | 141.5 | 79.8 | 5.5 | 0.3 | 3.6 | 27.4 |
| 28 | 3441.9 | 11.0 | 3.2 | 15.6 | 350.1 | 136.1 | 26.6 | 0.6 | 5.8 | 5.5 |
| 29 | 2489.5 | 111.5 | 2.0 | 16.4 | 280.8 | 141.8 | 10.3 | 0.5 | 101.5 | 101.3 |
| 30 | 1583.4 | 37.8 | 1.3 | 28.7 | 289.4 | 51.9 | 29.2 | 0.7 | 83.2 | 27.1 |
| 31 | 3708.2 | 26.4 | 4.1 | 37.3 | 417.6 | 178.2 | 31.7 | 0.2 | 91.0 | 33.9 |
| 32 | 1216.9 | 46.3 | 3.2 | 9.8 | 577.3 | 69.5 | 7.7 | 0.3 | 14.4 | 79.0 |
| 33 | 3493.2 | 12.6 | 11.4 | 17.5 | 417.4 | 159.5 | 7.3 | 1.3 | 12.3 | 51.1 |
| 34 | 1980.3 | 40.2 | 2.2 | 7.5 | 617.8 | 69.1 | 14.4 | 0.7 | 6.5 | 17.4 |
| 35 | 3867.8 | 17.6 | 3.2 | 14.4 | 68.5 | 118.9 | 31.9 | 0.8 | 5.8 | 29.7 |

**Table A.3:** Randomly drawn hierarchical parameter values used in the third Nrf2 simulation study.

| Cell ID | Stimulated condition | Basal condition |
|---|---|---|
| 1 | 85 | 64 |
| 2 | 64 | 100 |
| 3 | 97 | 95 |
| 4 | 98 | 100 |
| 5 | 27 | 100 |
| 6 | 97 | 59 |
| 7 | 97 | 96 |
| 8 | 97 | 46 |
| 9 | 97 | 36 |
| 10 | 67 | 73 |
| 11 | 59 | 73 |
| 12 | 92 | 46 |
| 13 | 78 | 26 |
| 14 | 55 | 37 |
| 15 | 16 | 52 |
| 16 | 92 | 51 |
| 17 | 4 | 52 |
| 18 | 54 | 61 |
| 19 | 75 | 9 |
| 20 | 80 | 60 |
| 21 | 85 | 63 |
| 22 | 80 | 10 |
| 23 | 91 | 61 |
| 24 | 25 | 21 |
| 25 | 9 | 12 |
| 26 | 68 | 4 |
| 27 | 40 | 94 |
| 28 | 72 | 42 |
| 29 | 41 | 15 |
| 30 | 95 | 73 |
| 31 | 100 | 57 |
| 32 | 55 | 47 |
| 33 | 56 | 64 |
| 34 | 36 | 28 |
| 35 | 97 | 29 |
| 36 | 12 | - |
| Overall | 66.5 | 53.0 |

**Table A.4:** Percentage of complex eigenvalues in each cell, from the 100 hierarchical parameter values, under the basal and stimulated conditions.

## A.2 Part II

| Experiment | $\tilde{\alpha}_0^{(k)}$ | $\tilde{\alpha}_1^{(k)}$ | $\tilde{k}_{ON}^{(k)}$ | $\tilde{k}_{OFF}^{(k)}$ | $\kappa^{(k)}$ | $\mu_\epsilon^{(k)}$ | $\sigma_\epsilon^{(k)}$ |
|---|---|---|---|---|---|---|---|
| 1 | 47.86 | 449.24 | 0.28 | 6.59 | 116.26 | 775.26 | 331.27 |
| 2 | 63.78 | 465.73 | 0.19 | 6.15 | 84.93 | 800.32 | 309.99 |
| 3 | 47.40 | 445.79 | 0.22 | 6.15 | 106.58 | 966.62 | 395.63 |
| 4 | 57.74 | 498.31 | 0.23 | 5.90 | 92.17 | 892.34 | 304.54 |

**Table A.5:** Randomly drawn hierarchical parameter values used in the first simulation study for the two-state switch gene model.

| Experiment | $\tilde{\alpha}_0^{(k)}$ | $\tilde{\alpha}_1^{(k)}$ | $\tilde{k}_{ON}^{(k)}$ | $\tilde{k}_{OFF}^{(k)}$ | $\kappa^{(k)}$ | $\mu_\epsilon^{(k)}$ | $\sigma_\epsilon^{(k)}$ |
|---|---|---|---|---|---|---|---|
| 1 | 42.34 | 463.82 | 0.19 | 7.48 | 21.82 | 775.26 | 331.27 |
| 2 | 53.75 | 567.93 | 0.19 | 6.07 | 15.92 | 800.32 | 309.99 |
| 3 | 56.81 | 457.53 | 0.21 | 6.19 | 21.60 | 966.62 | 395.63 |
| 4 | 51.31 | 496.38 | 0.24 | 6.62 | 24.78 | 892.34 | 304.54 |

**Table A.6:** Randomly drawn hierarchical parameter values used in the second simulation study for the two-state switch gene model.

| Experiment | $\tilde{\alpha}_0^{(k)}$ | $\tilde{\alpha}_1^{(k)}$ | $\tilde{k}_{ON}^{(k)}$ | $\tilde{k}_{OFF}^{(k)}$ | $\kappa^{(k)}$ | $\mu_\epsilon^{(k)}$ | $\sigma_\epsilon^{(k)}$ |
|---|---|---|---|---|---|---|---|
| 1 | 43.59 | 511.77 | 0.24 | 6.30 | 20.58 | 775.26 | 331.27 |
| 2 | 42.06 | 570.16 | 0.20 | 6.41 | 17.61 | 800.32 | 309.99 |
| 3 | 49.58 | 438.01 | 0.23 | 6.51 | 19.31 | 966.62 | 395.63 |
| 4 | 41.44 | 488.60 | 0.19 | 6.25 | 22.71 | 892.34 | 304.54 |

**Table A.7:** Randomly drawn hierarchical parameter values used in the third simulation study for the two-state switch gene model.

| *Tetracycline* | Experiment | | | | |
|---|---|---|---|---|---|
| ng/ml | 1 | 2 | 3 | 4 | Average |
| 0 | 2021.7 | 1125.0 | 1226.9 | 1514.5 | 1472.0 |
| 5 | 1871.2 | 2250.2 | 1729.3 | 2513.7 | 2091.1 |
| 10 | 3823.4 | 2702.0 | 2357.0 | 3402.0 | 3071.1 |

**Table A.8:** Average of the observations in each experiment, $\langle Y^{(k)} \rangle$, for all experimental conditions. The last column represents the average across the four experiments.

| Parameter | Experiment 1 | | Experiment 2 | | Experiment 3 | | Experiment 4 | |
|---|---|---|---|---|---|---|---|---|
| | LB | UB | LB | UB | LB | UB | LB | UB |
| $\tilde{\alpha}_0^{(k)}$ | 13.10 | 19.18 | 3.57 | 8.26 | 6.61 | 10.30 | 18.64 | 26.50 |
| $\tilde{\alpha}_1^{(k)}$ | 2545.85 | 3224.16 | 4864.22 | 6745.57 | 4072.51 | 5522.62 | 2738.33 | 3879.20 |
| $\tilde{k}_{ON}^{(k)}$ | 0.50 | 0.61 | 0.20 | 0.25 | 0.09 | 0.12 | 0.23 | 0.28 |
| $\tilde{k}_{OFF}^{(k)}$ | 44.36 | 63.54 | 20.14 | 25.71 | 12.49 | 16.06 | 20.66 | 25.92 |
| $\kappa^{(k)}$ | 22.42 | 32.48 | 4.36 | 6.33 | 5.33 | 6.72 | 8.47 | 12.44 |
| $\mu_\epsilon^{(k)}$ | 767.30 | 783.74 | 789.42 | 808.60 | 958.70 | 976.21 | 886.17 | 897.98 |
| $\sigma_\epsilon^{(k)}$ | 325.69 | 336.72 | 291.66 | 303.52 | 359.70 | 370.01 | 299.76 | 307.47 |
| $\frac{\alpha_0^{(k)}}{\alpha_1^{(k)}} \times 10^2$ | 0.46 | 0.64 | 0.06 | 0.17 | 0.14 | 0.22 | 0.61 | 0.75 |
| $\mu_P^{(k)} \times 10^2$ | 0.87 | 1.20 | 0.88 | 1.08 | 0.67 | 0.81 | 0.99 | 1.22 |
| $\frac{1}{\tilde{k}_{ON}^{(k)}} \times 10^2$ | 162.74 | 197.77 | 397.03 | 506.22 | 851.24 | 1069.14 | 349.82 | 424.30 |
| $\frac{1}{\tilde{k}_{OFF}^{(k)}} \times 10^2$ | 1.55 | 2.23 | 3.86 | 4.92 | 6.21 | 7.98 | 3.84 | 4.82 |
| $\mu_X^{(k)}$ | 37.51 | 54.48 | 51.17 | 75.35 | 38.03 | 48.79 | 49.24 | 72.54 |
| $\mu_Y^{(k)}$ | 1995.53 | 2051.64 | 1110.78 | 1141.46 | 1215.77 | 1238.95 | 1500.13 | 1530.05 |

**Table A.9:** 0.95 level HPD CIs for the hierarchical parameters, for the non-stimulated condition.

| Parameter | Experiment 1 | | Experiment 2 | | Experiment 3 | | Experiment 4 | |
|---|---|---|---|---|---|---|---|---|
| | LB | UB | LB | UB | LB | UB | LB | UB |
| $\tilde{\alpha}_0^{(k)}$ | 2.49 | 9.76 | 1.50 | 4.21 | 17.31 | 24.72 | 8.97 | 16.41 |
| $\tilde{\alpha}_1^{(k)}$ | 4144.27 | 4747.91 | 3496.21 | 3686.24 | 3677.01 | 4205.64 | 3922.07 | 4267.06 |
| $\tilde{k}_{ON}^{(k)}$ | 0.65 | 0.75 | 0.71 | 0.77 | 0.29 | 0.33 | 0.75 | 0.83 |
| $\tilde{k}_{OFF}^{(k)}$ | 12.80 | 16.60 | 8.20 | 9.25 | 12.00 | 16.40 | 12.26 | 14.95 |
| $\kappa^{(k)}$ | 4.86 | 5.61 | 5.00 | 5.20 | 6.04 | 7.95 | 6.12 | 7.58 |
| $\mu_\epsilon^{(k)}$ | 766.16 | 782.63 | 786.33 | 804.91 | 957.47 | 976.90 | 886.51 | 898.36 |
| $\sigma_\epsilon^{(k)}$ | 321.12 | 331.99 | 297.23 | 309.96 | 371.89 | 383.54 | 297.65 | 305.54 |
| $\frac{\alpha_0^{(k)}}{\alpha_1^{(k)}} \times 10^2$ | 0.05 | 0.23 | 0.04 | 0.12 | 0.42 | 0.62 | 0.22 | 0.40 |
| $\mu_P^{(k)} \times 10^2$ | 4.12 | 4.99 | 7.48 | 8.09 | 1.88 | 2.42 | 5.05 | 5.94 |
| $\frac{1}{\tilde{k}_{ON}^{(k)}} \times 10^2$ | 132.50 | 152.98 | 129.86 | 141.22 | 300.55 | 345.63 | 119.83 | 133.43 |
| $\frac{1}{\tilde{k}_{OFF}^{(k)}} \times 10^2$ | 6.01 | 7.79 | 10.76 | 12.14 | 6.01 | 8.21 | 6.65 | 8.11 |
| $\mu_X^{(k)}$ | 192.94 | 224.91 | 275.35 | 293.03 | 93.25 | 123.58 | 213.06 | 265.48 |
| $\mu_Y^{(k)}$ | 1836.92 | 1901.56 | 2213.79 | 2281.58 | 1710.53 | 1747.34 | 2480.27 | 2541.93 |

**Table A.10:** 0.95 level HPD CIs for the hierarchical parameters, under the induction by *tetracycline* at 5 ng/ml.

| Parameter | Experiment 1 | | Experiment 2 | | Experiment 3 | | Experiment 4 | |
|---|---|---|---|---|---|---|---|---|
| | LB | UB | LB | UB | LB | UB | LB | UB |
| $\tilde{\alpha}_0^{(k)}$ | 0.00 | 7.71 | 0.00 | 2.23 | 16.12 | 26.57 | 0.00 | 1.70 |
| $\tilde{\alpha}_1^{(k)}$ | 2223.72 | 2404.12 | 2192.18 | 2419.42 | 3689.72 | 3941.69 | 1802.03 | 2022.14 |
| $\tilde{k}_{ON}^{(k)}$ | 1.05 | 1.17 | 0.77 | 0.85 | 0.67 | 0.77 | 1.13 | 1.22 |
| $\tilde{k}_{OFF}^{(k)}$ | 4.66 | 5.40 | 3.33 | 4.20 | 15.10 | 21.29 | 8.88 | 11.05 |
| $\kappa^{(k)}$ | 6.92 | 7.42 | 4.25 | 5.09 | 7.13 | 9.66 | 12.21 | 13.06 |
| $\mu_\epsilon^{(k)}$ | 769.53 | 786.58 | 780.83 | 800.02 | 957.32 | 977.87 | 885.94 | 897.62 |
| $\sigma_\epsilon^{(k)}$ | 322.26 | 333.30 | 304.38 | 317.95 | 379.61 | 392.83 | 298.66 | 306.73 |
| $\frac{\alpha_0^{(k)}}{\alpha_1^{(k)}} \times 10^2$ | 0.00 | 0.34 | 0.00 | 0.09 | 0.42 | 0.72 | 0.00 | 0.09 |
| $\mu_P^{(k)} \times 10^2$ | 17.47 | 19.03 | 16.48 | 18.92 | 3.33 | 4.42 | 9.71 | 11.33 |
| $\frac{1}{\tilde{k}_{ON}^{(k)}} \times 10^2$ | 85.48 | 95.09 | 117.30 | 130.06 | 129.28 | 147.54 | 82.24 | 88.52 |
| $\frac{1}{\tilde{k}_{OFF}^{(k)}} \times 10^2$ | 18.52 | 21.45 | 23.64 | 29.79 | 4.64 | 6.53 | 8.98 | 11.17 |
| $\mu_X^{(k)}$ | 404.40 | 443.46 | 371.80 | 455.53 | 142.84 | 193.83 | 190.50 | 206.49 |
| $\mu_Y^{(k)}$ | 3754.91 | 3882.85 | 2656.61 | 2746.80 | 2332.38 | 2379.31 | 3358.54 | 3440.29 |

**Table A.11:** 0.95 level HPD CIs for the hierarchical parameters, under the induction by *tetracycline* at 10 ng/ml.

# Appendix B

# Figures

## B.1 Part I

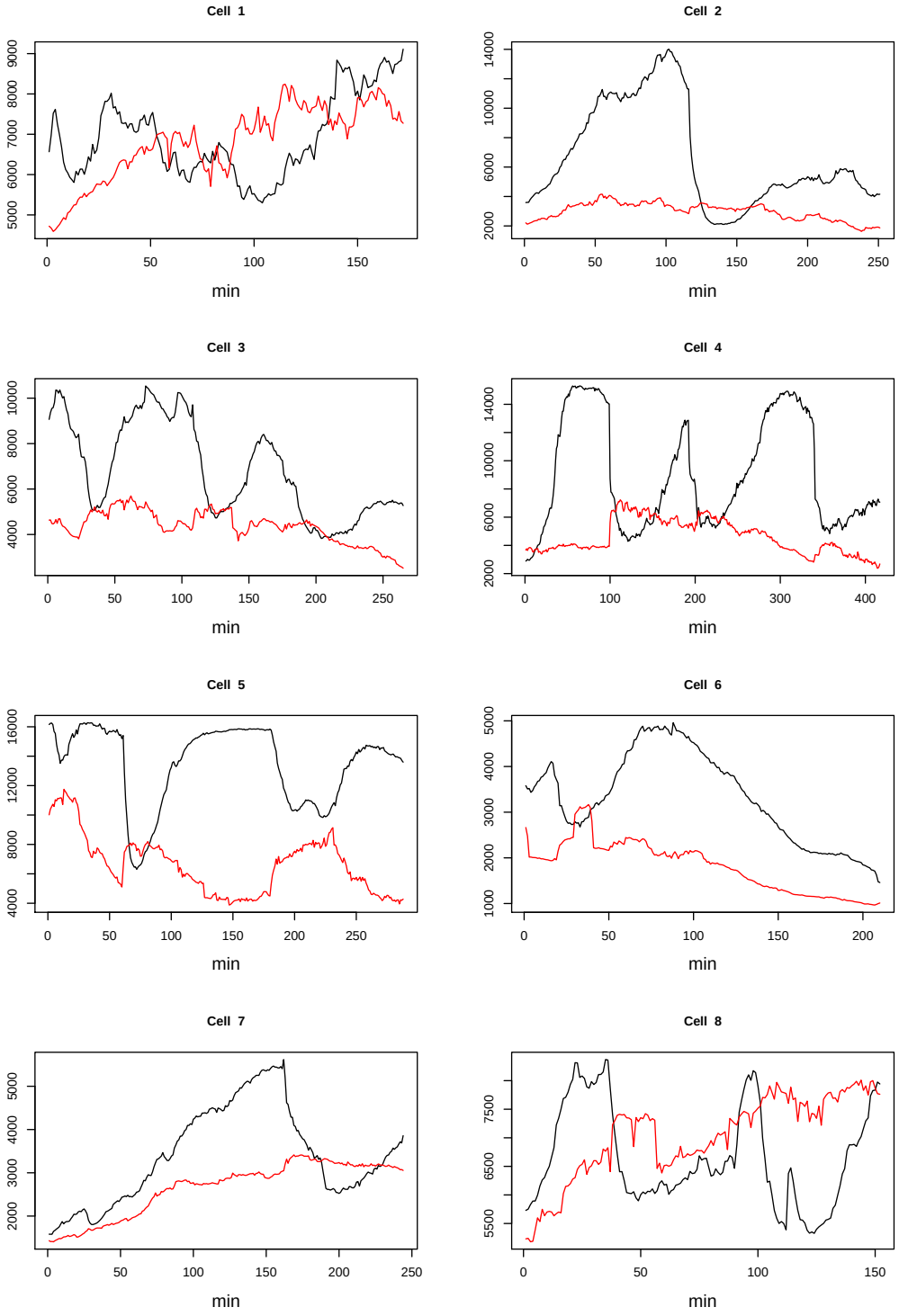


**Figure B.1:** Experimental data observed, under the basal condition, for $y_t^N$ (in black) and $y_t^C$ (in red) for cells 1 to 8.

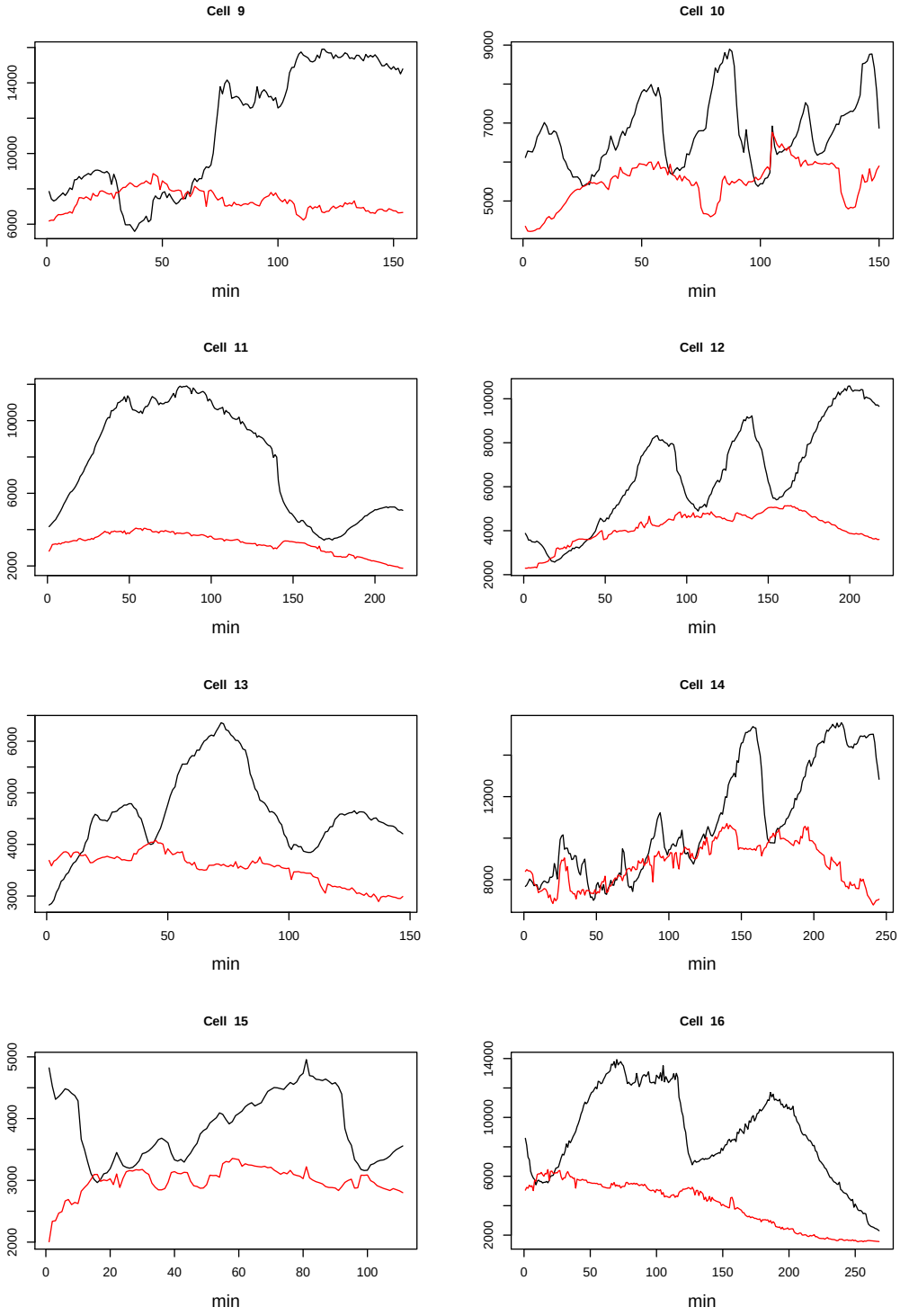


**Figure B.2:** Experimental data observed, under the basal condition, for $y_t^N$ (in black) and $y_t^C$ (in red) for cells 9 to 16.

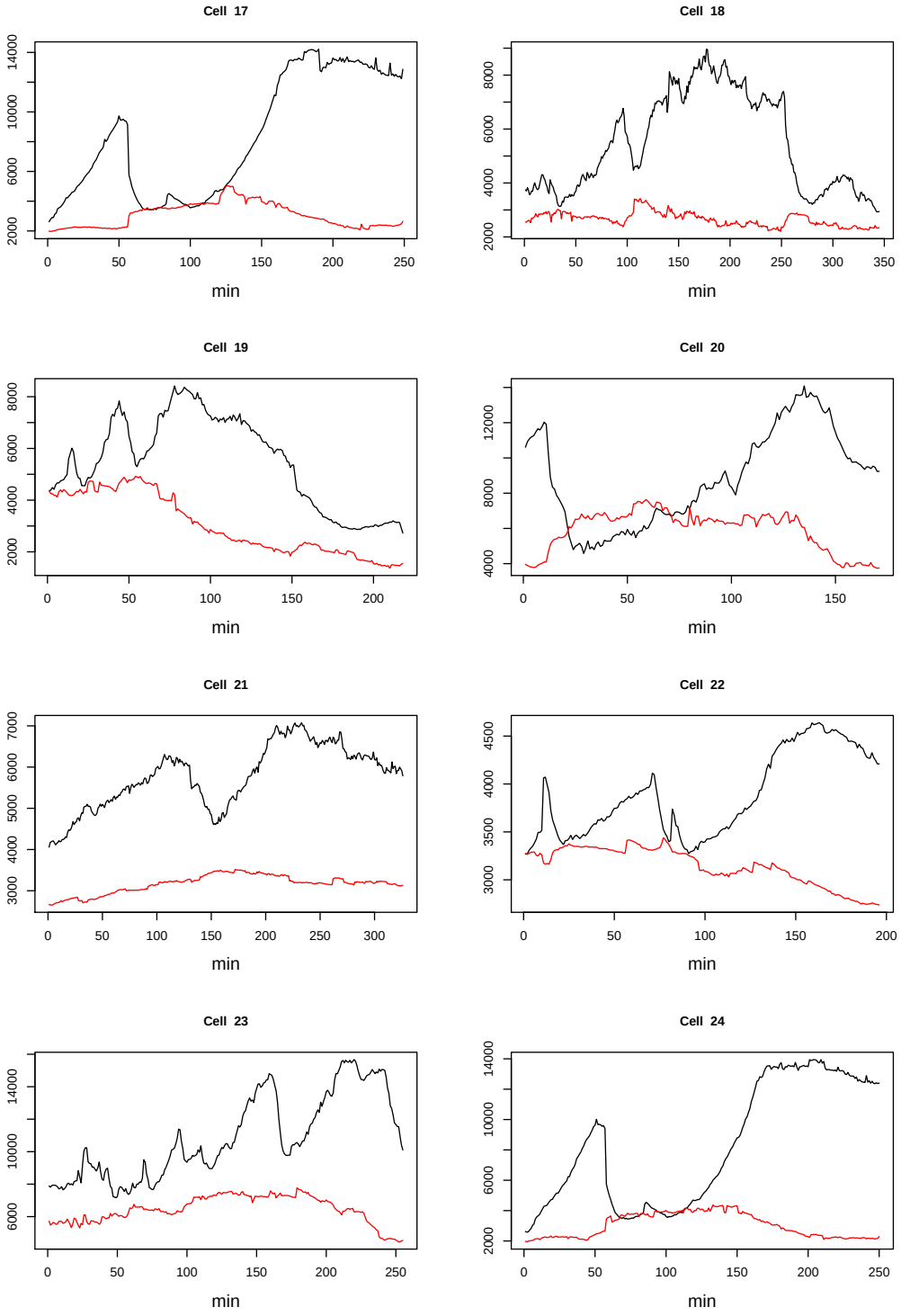


**Figure B.3:** Experimental data observed, under the basal condition, for $y_t^N$ (in black) and $y_t^C$ (in red) for cells 17 to 24.

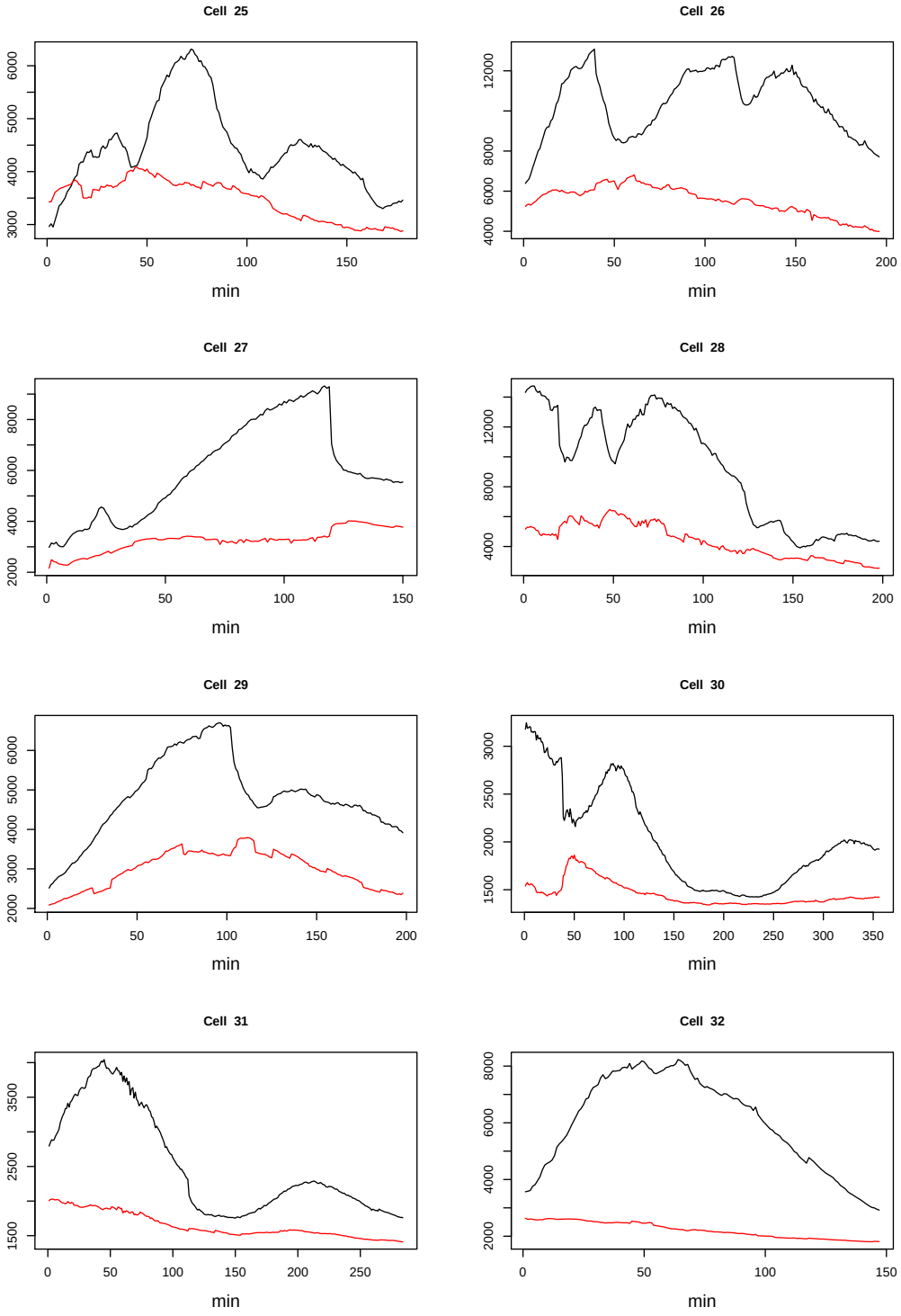


**Figure B.4:** Experimental data observed, under the basal condition, for $y_t^N$ (in black) and $y_t^C$ (in red) for cells 25 to 32.

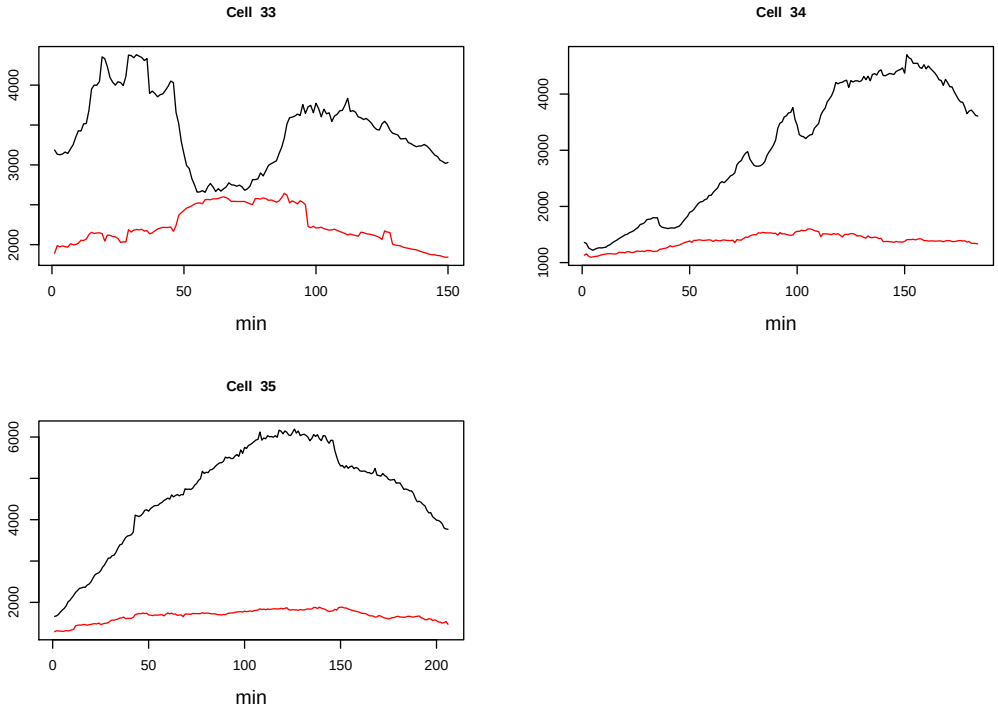


**Figure B.5:** Experimental data observed, under the basal condition, for $y_t^N$ (in black) and $y_t^C$ (in red) for cells 33 to 35.

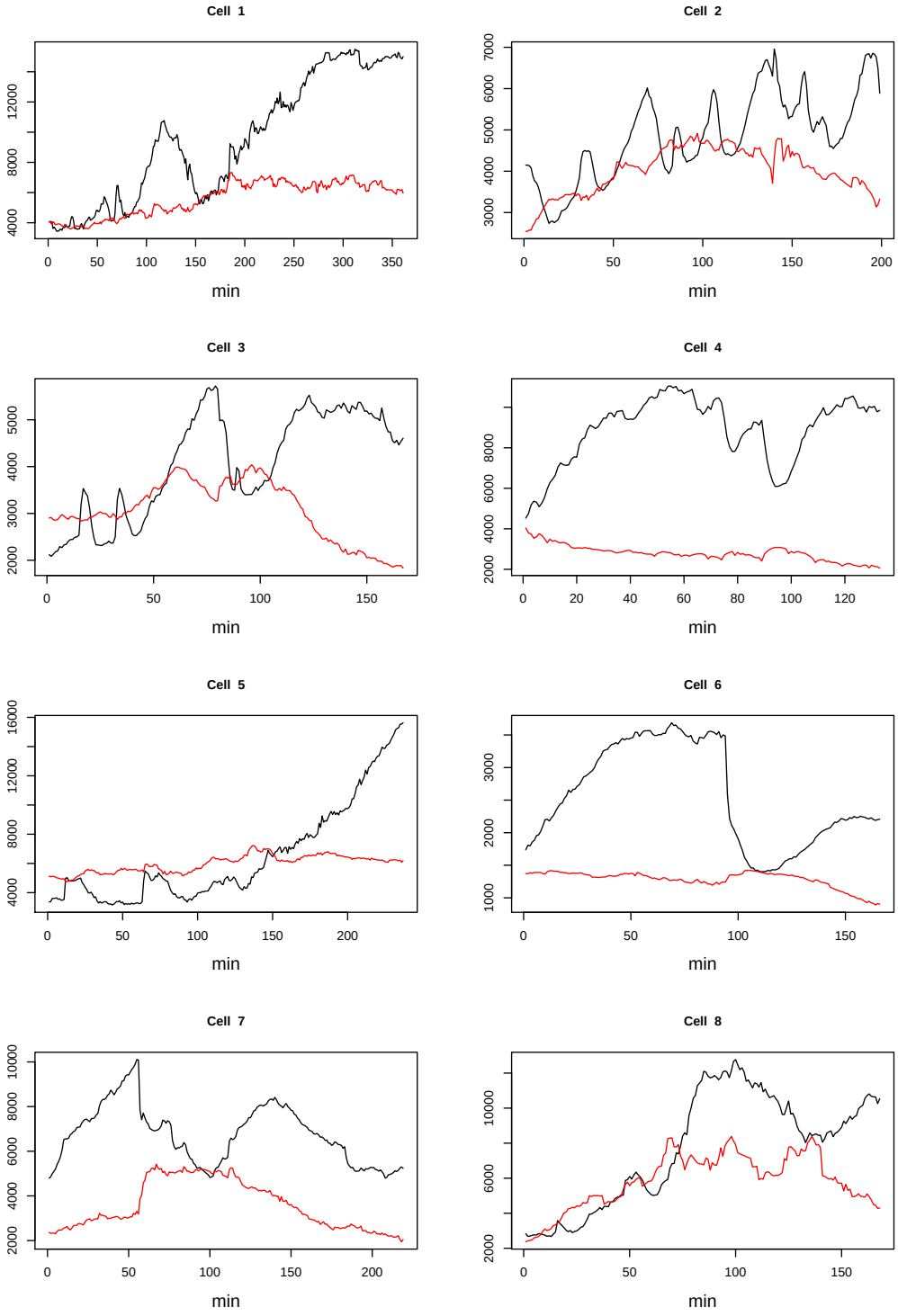


**Figure B.6:** Experimental data observed, under the induction of SFN stimulant, for $y_t^N$ (in black) and $y_t^C$ (in red) for cells 1 to 8.

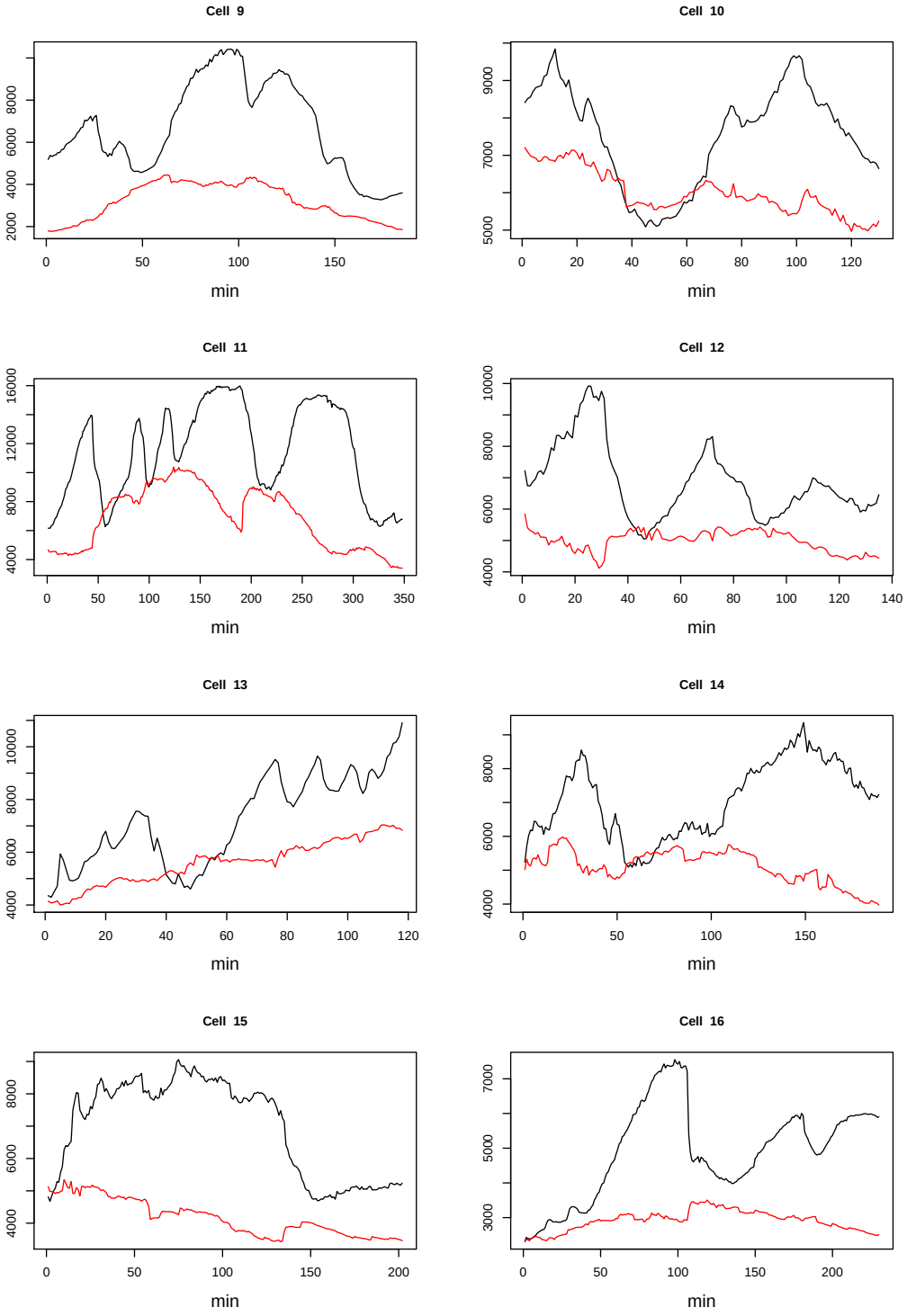


**Figure B.7:** Experimental data observed, under the induction of SFN stimulant, for $y_t^N$ (in black) and $y_t^C$ (in red) for cells 9 to 16.

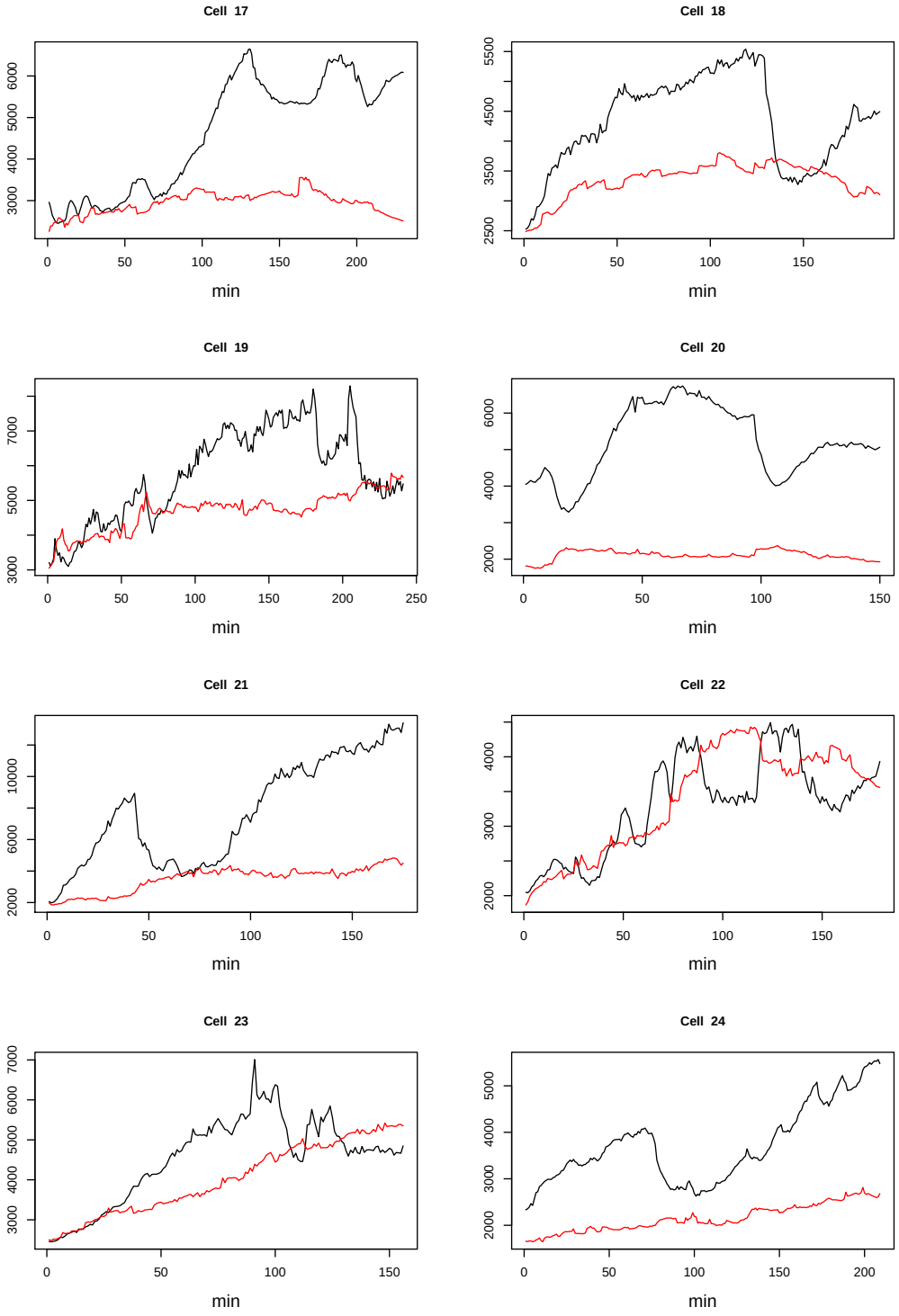


**Figure B.8:** Experimental data observed, under the induction of SFN stimulant, for $y_t^N$ (in black) and $y_t^C$ (in red) for cells 17 to 24.

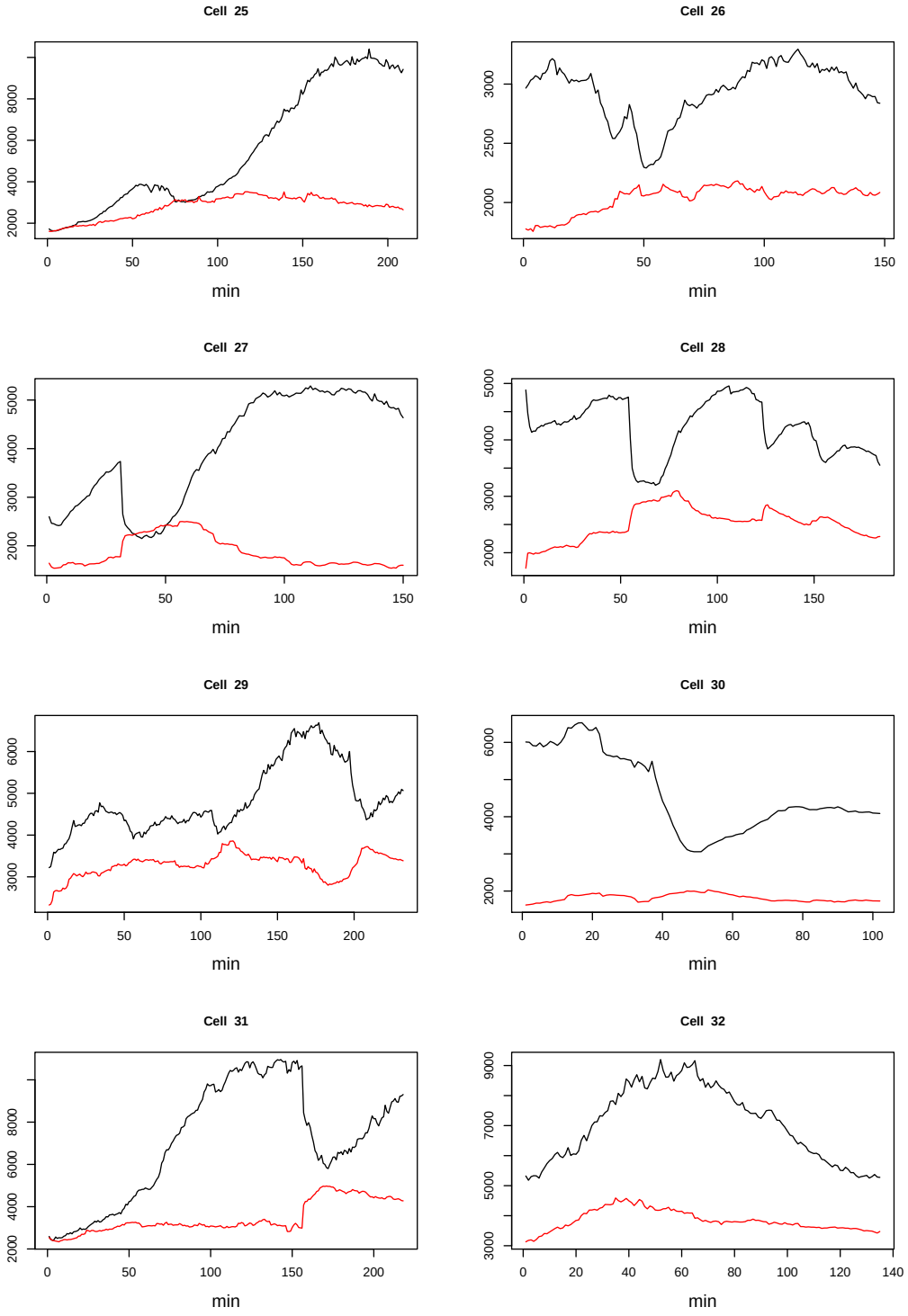


**Figure B.9:** Experimental data observed, under the induction of SFN stimulant, for $y_t^N$ (in black) and $y_t^C$ (in red) for cells 25 to 32.

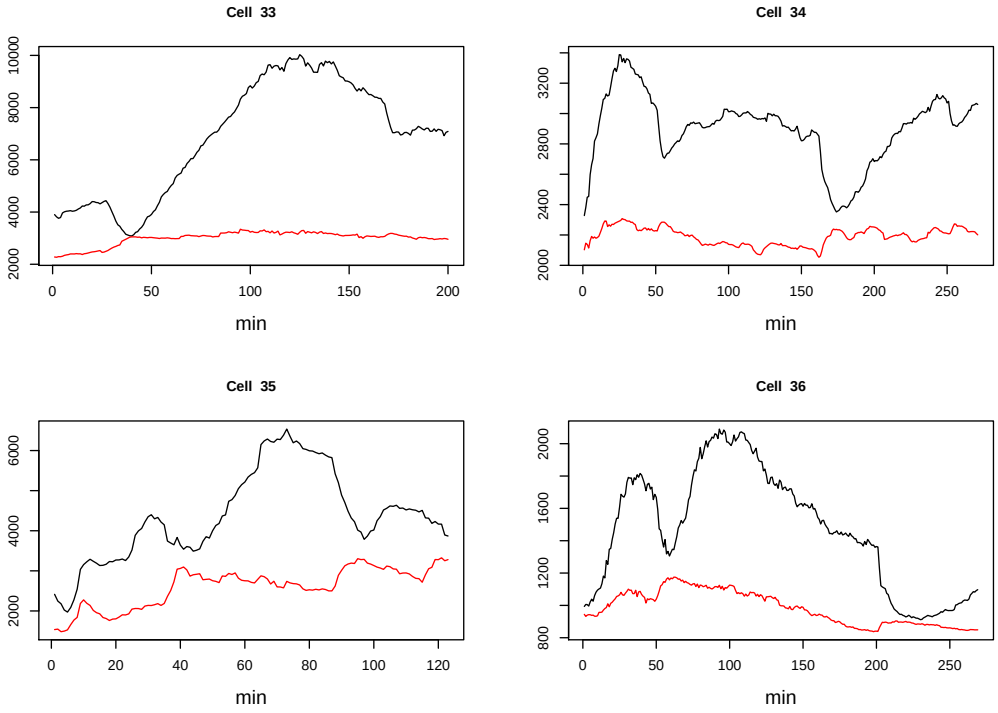


**Figure B.10:** Experimental data observed, under the induction of SFN stimulant, for $y_t^N$ (in black) and $y_t^C$ (in red) for cells 33 to 36.

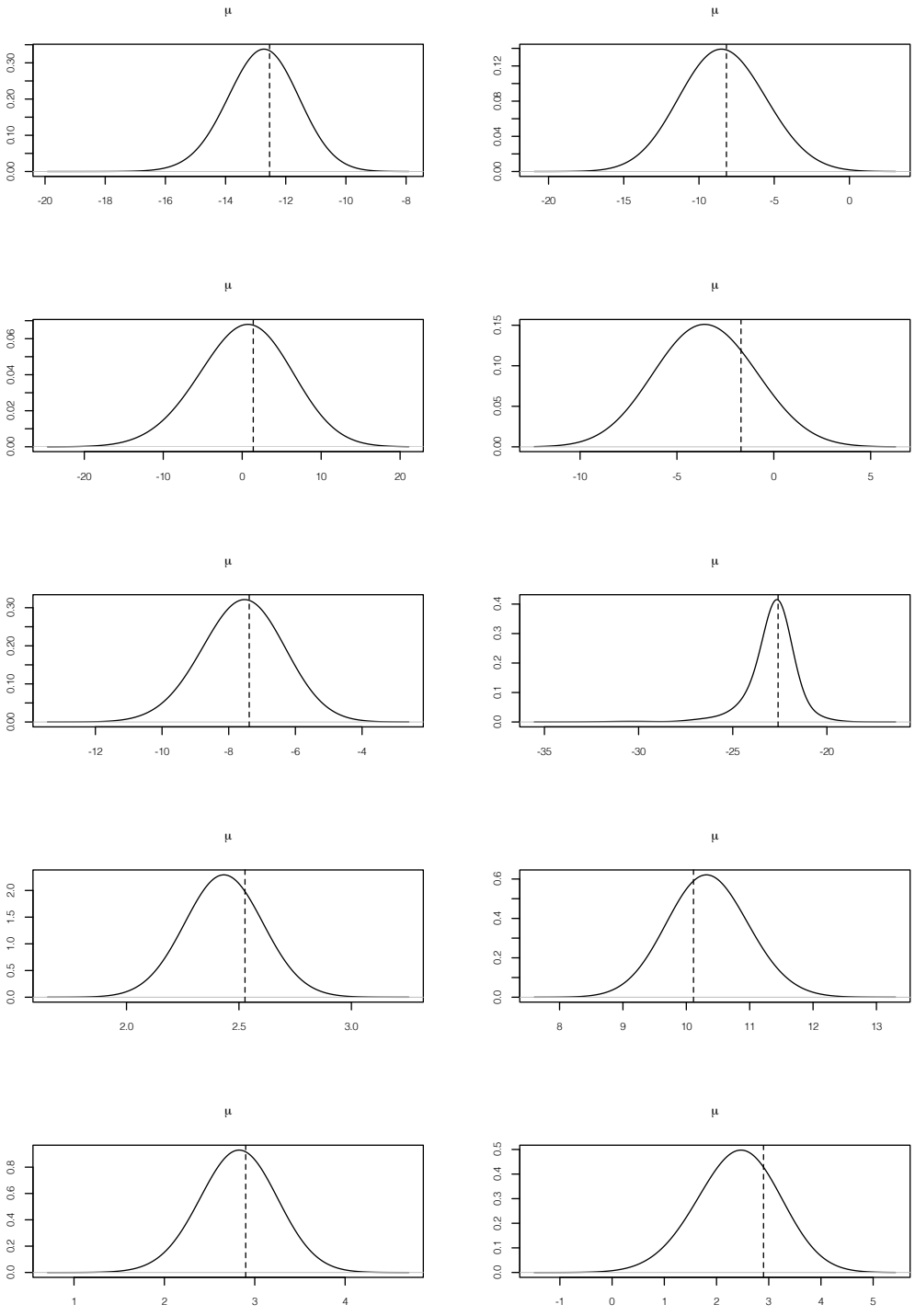


**Figure B.11:** Posterior densities for the hypermean parameters from the first simulation study. The vertical dashed line denotes the real value used in the simulation. Figures refer, from left to right and top to bottom, to the hypermean parameters of $k_d^{(i)}, k_a^{(i)}, K_a^{(i)}, \mu_\tau^{(i)}, \gamma^{(i)}, \delta^{(i)}, c^{(i)}, \kappa_N^{(i)}, \sigma_N^{(i)}$ and $\sigma_C^{(i)}$.

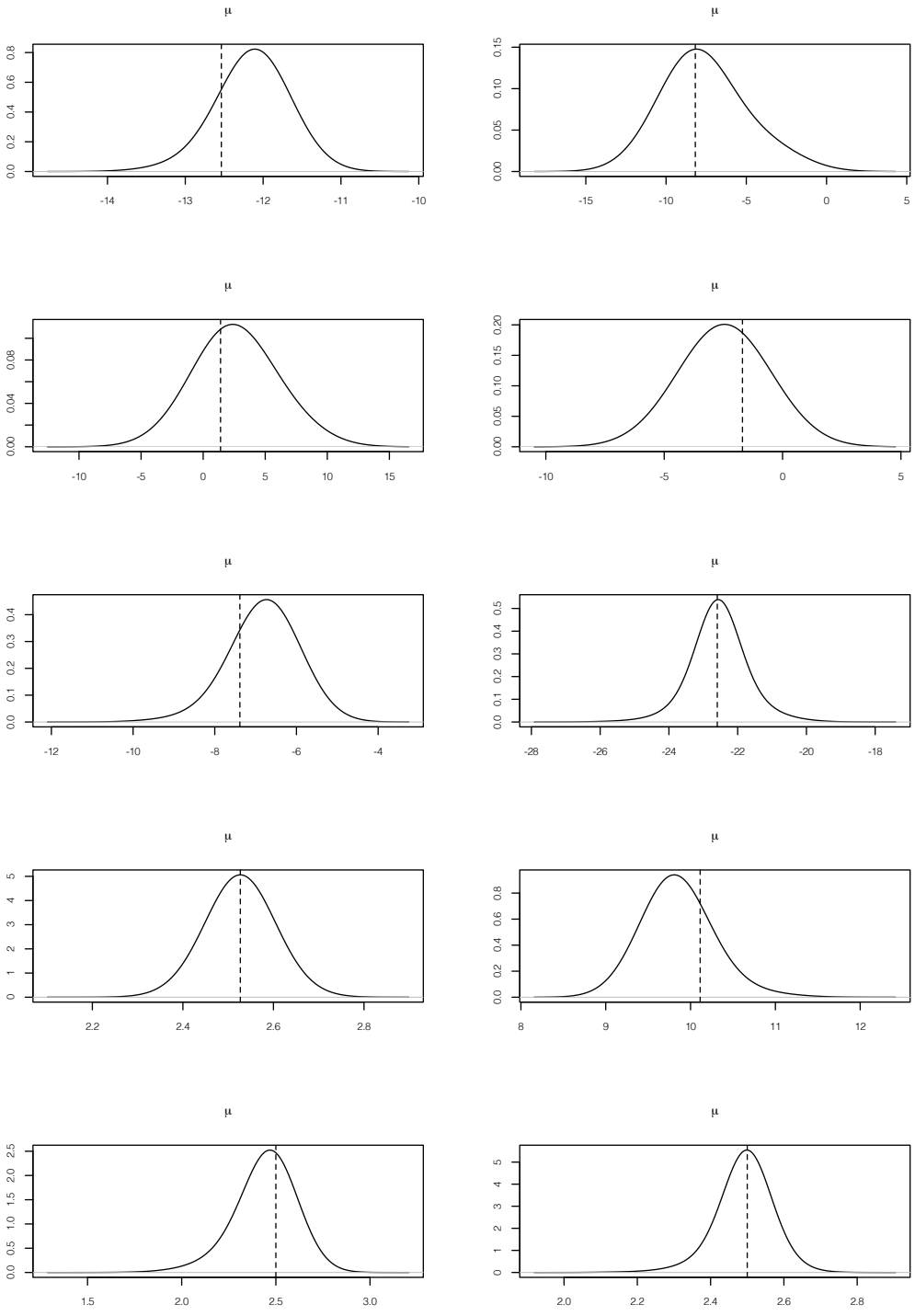


**Figure B.12:** Posterior densities for the hypermean parameters from the second simulation study. The vertical dashed line denotes the real value used in the simulation. Figures refer, from left to right and top to bottom, to the hypermean parameters of $k_d^{(i)}, k_a^{(i)}, K_a^{(i)}, \mu_\tau^{(i)}, \gamma^{(i)}, \delta^{(i)}, c^{(i)}, \kappa_N^{(i)}, \sigma_N^{(i)}$ and $\sigma_C^{(i)}$.

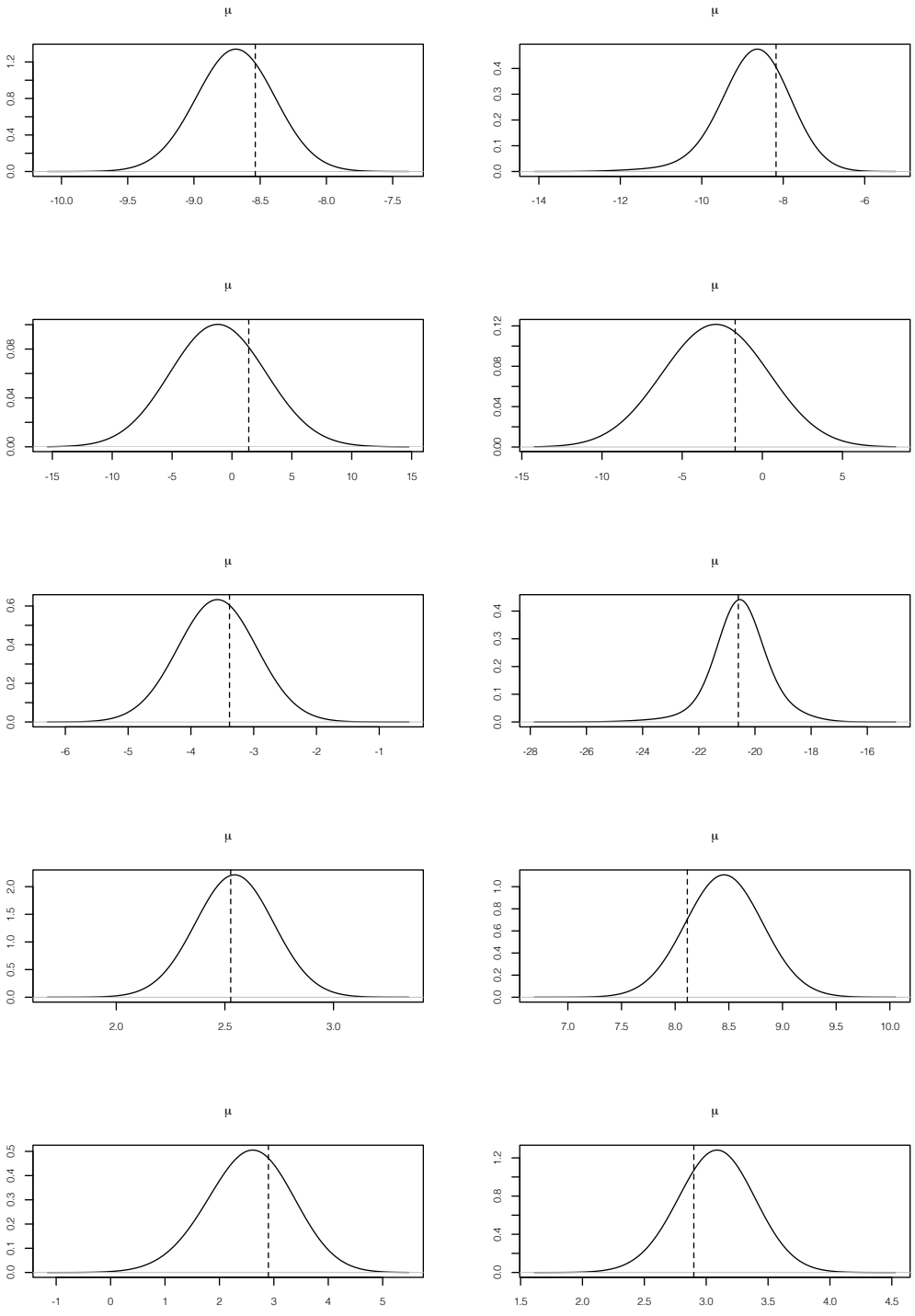


**Figure B.13:** Posterior densities for the hypermean parameters from the third simulation study. The vertical dashed line denotes the real value used in the simulation. Figures refer, from left to right and top to bottom, to the hypermean parameters of $k_d^{(i)}, k_a^{(i)}, K_a^{(i)}, \mu_\tau^{(i)}, \gamma^{(i)}, \delta^{(i)}, c^{(i)}, \kappa_N^{(i)}, \sigma_N^{(i)}$ and $\sigma_C^{(i)}$.

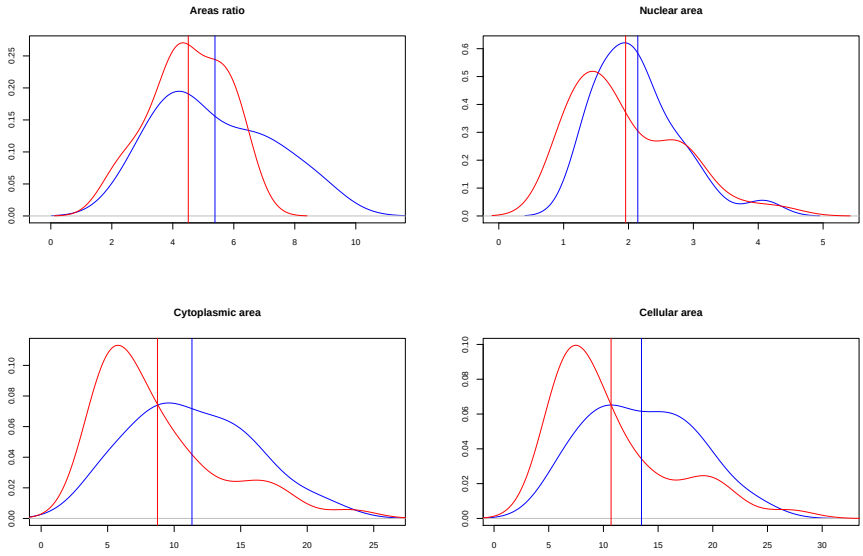


**Figure B.14:** Densities of the areas, computed at time t = 31, for the cells under the basal condition (in blue) and under the induction of a stimulant (in red). The vertical solid lines denote the mean values.

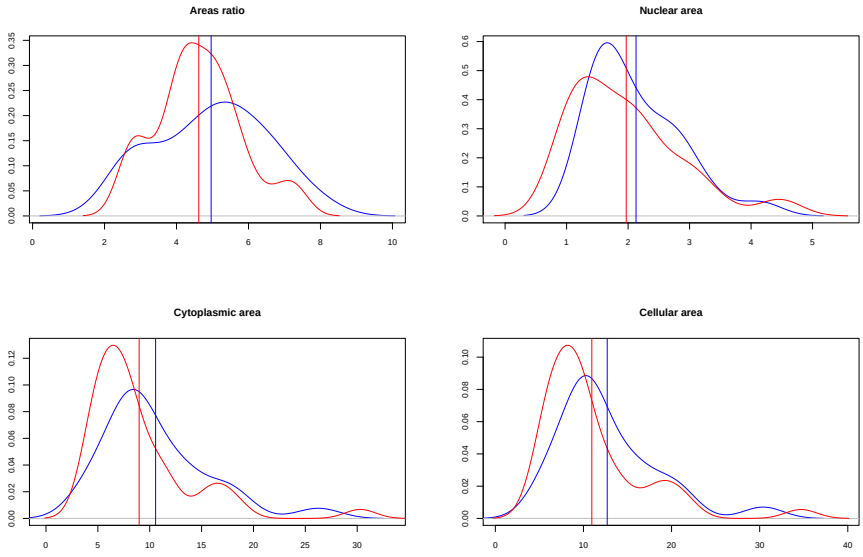


**Figure B.15:** Densities of the areas, computed at time t = 61, for the cells under the basal condition (in blue) and under the induction of a stimulant (in red). The vertical solid lines denote the mean values.

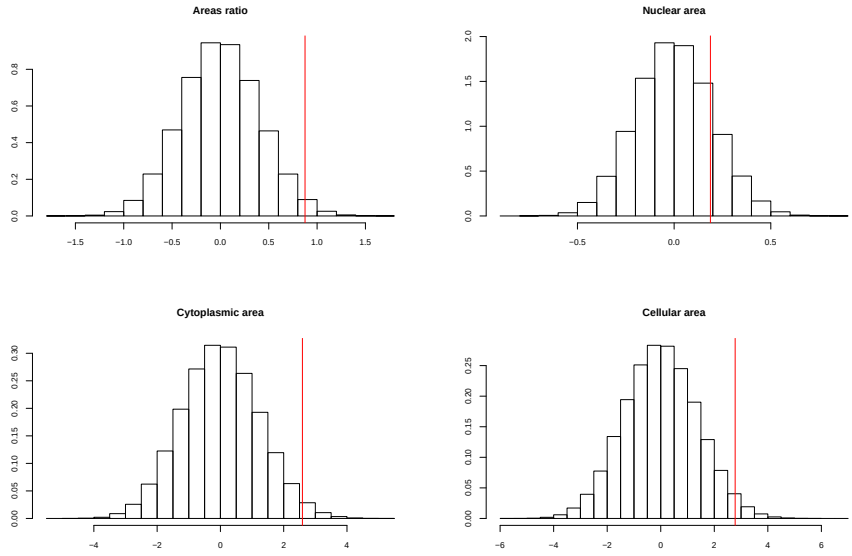


**Figure B.16:** Null distribution, for the mean difference, of permutation tests on the mean of the ratio of areas and of nuclear, cytoplasmic and cellular areas, at time 31. The vertical solid line represents the observed value of the statistic.

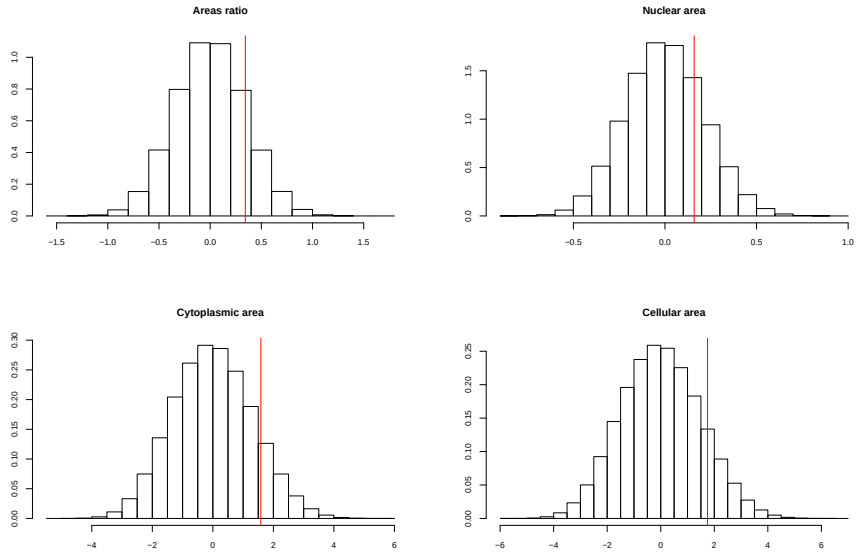


**Figure B.17:** Null distribution, for the mean difference, of permutation tests on the mean of the ratio of areas and of nuclear, cytoplasmic and cellular areas, at time 61. The vertical solid line represents the observed value of the statistic.

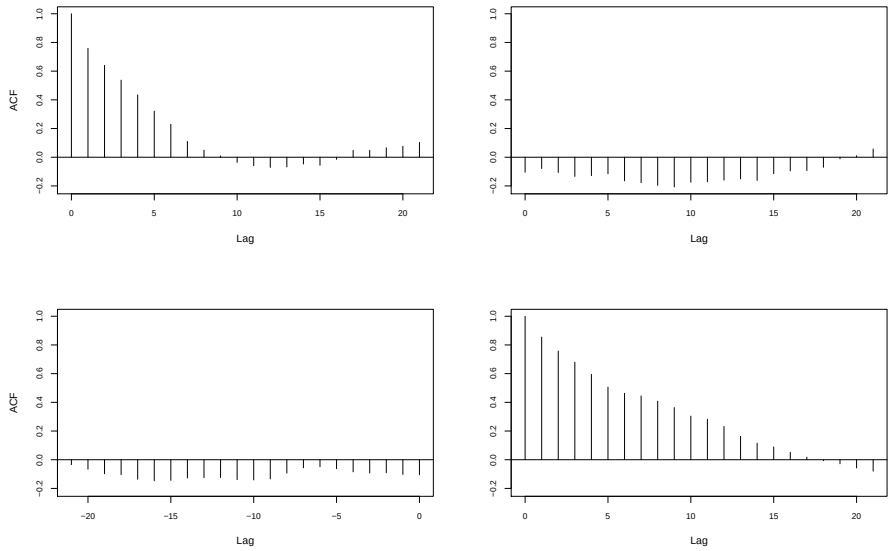


**Figure B.18:** Autocorrelation plot of the difference between repeated measurements, $y_t^I - y_t^{II} = \epsilon_t^{CT^I} - \epsilon_t^{CT^{II}}$, for the nuclear (top left) and cytoplasmic (bottom right) concentrations in cell 1.

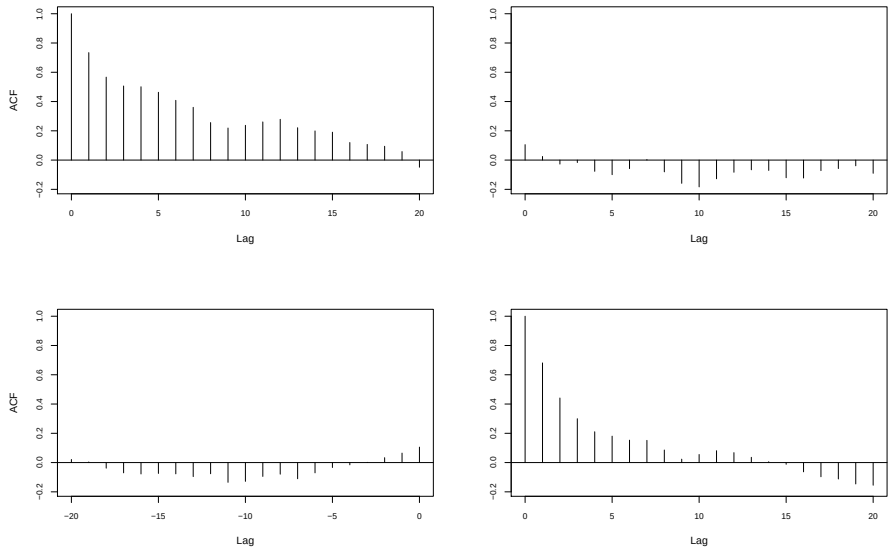


**Figure B.19:** Autocorrelation plot of the difference between repeated measurements, $y_t^I - y_t^{II} = \epsilon_t^{CT^I} - \epsilon_t^{CT^{II}}$, for the nuclear (top left) and cytoplasmic (bottom right) concentrations in cell 2.

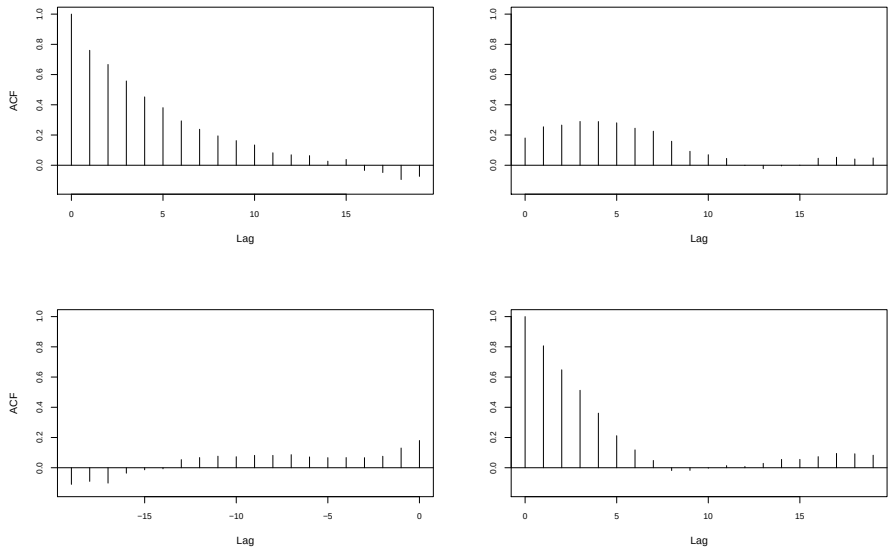


**Figure B.20:** Autocorrelation plot of the difference between repeated measurements, $y_t^I - y_t^{II} = \epsilon_t^{CT^I} - \epsilon_t^{CT^{II}}$, for the nuclear (top left) and cytoplasmic (bottom right) concentrations in cell 3.

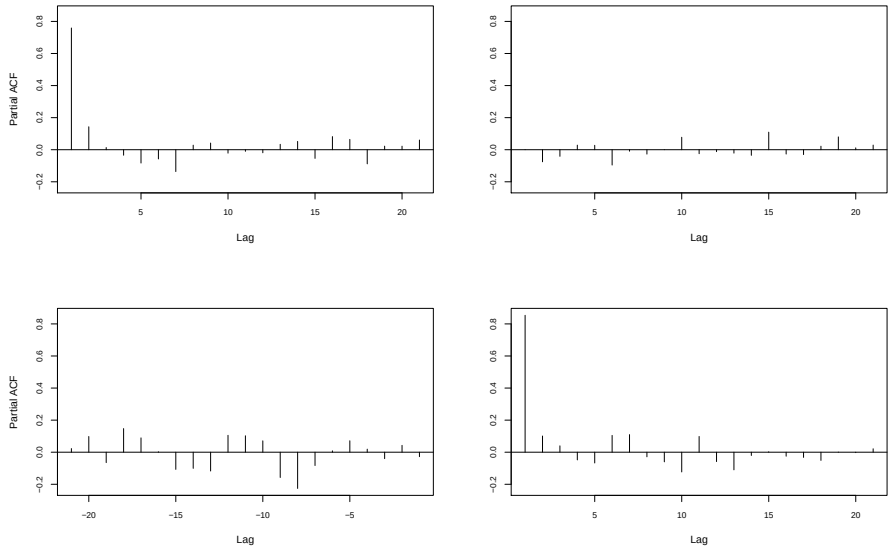


**Figure B.21:** Partial autocorrelation plot of the difference between repeated measurements, $y_t^I - y_t^{II} = \epsilon_t^{CT^I} - \epsilon_t^{CT^{II}}$, for the nuclear (top left) and cytoplasmic (bottom right) concentrations in cell 1.

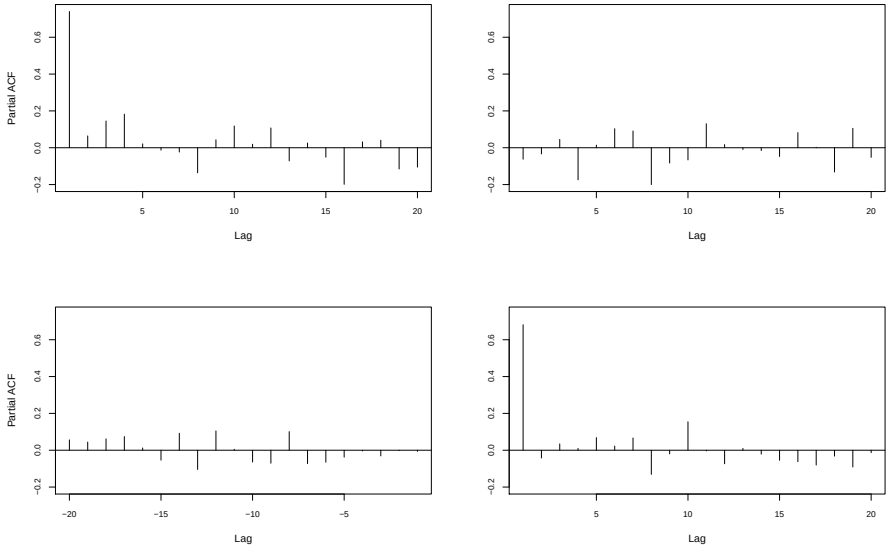


**Figure B.22:** Partial autocorrelation plot of the difference between repeated measurements, $y_t^I - y_t^{II} = \epsilon_t^{CT^I} - \epsilon_t^{CT^{II}}$, for the nuclear (top left) and cytoplasmic (bottom right) concentrations in cell 2.

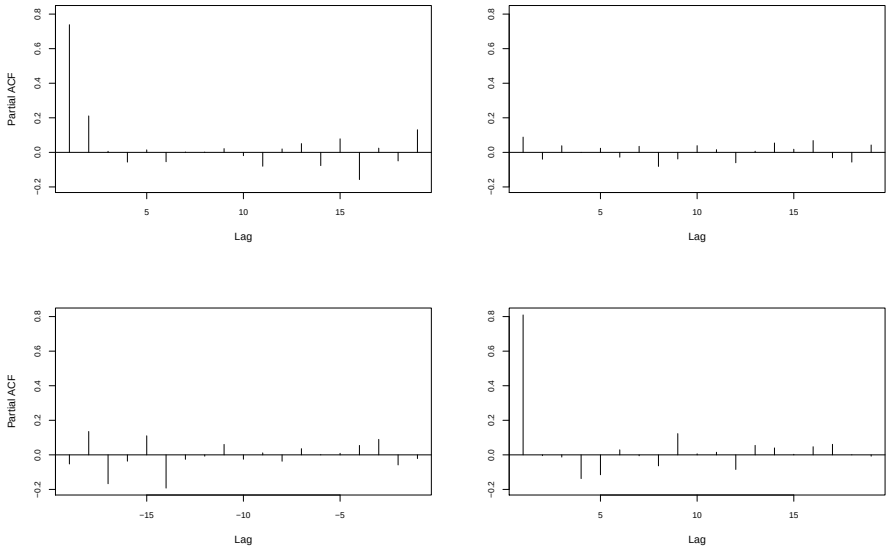


**Figure B.23:** Partial autocorrelation plot of the difference between repeated measurements, $y_t^I - y_t^{II} = \epsilon_t^{CT^I} - \epsilon_t^{CT^{II}}$, for the nuclear (top left) and cytoplasmic (bottom right) concentrations in cell 3.

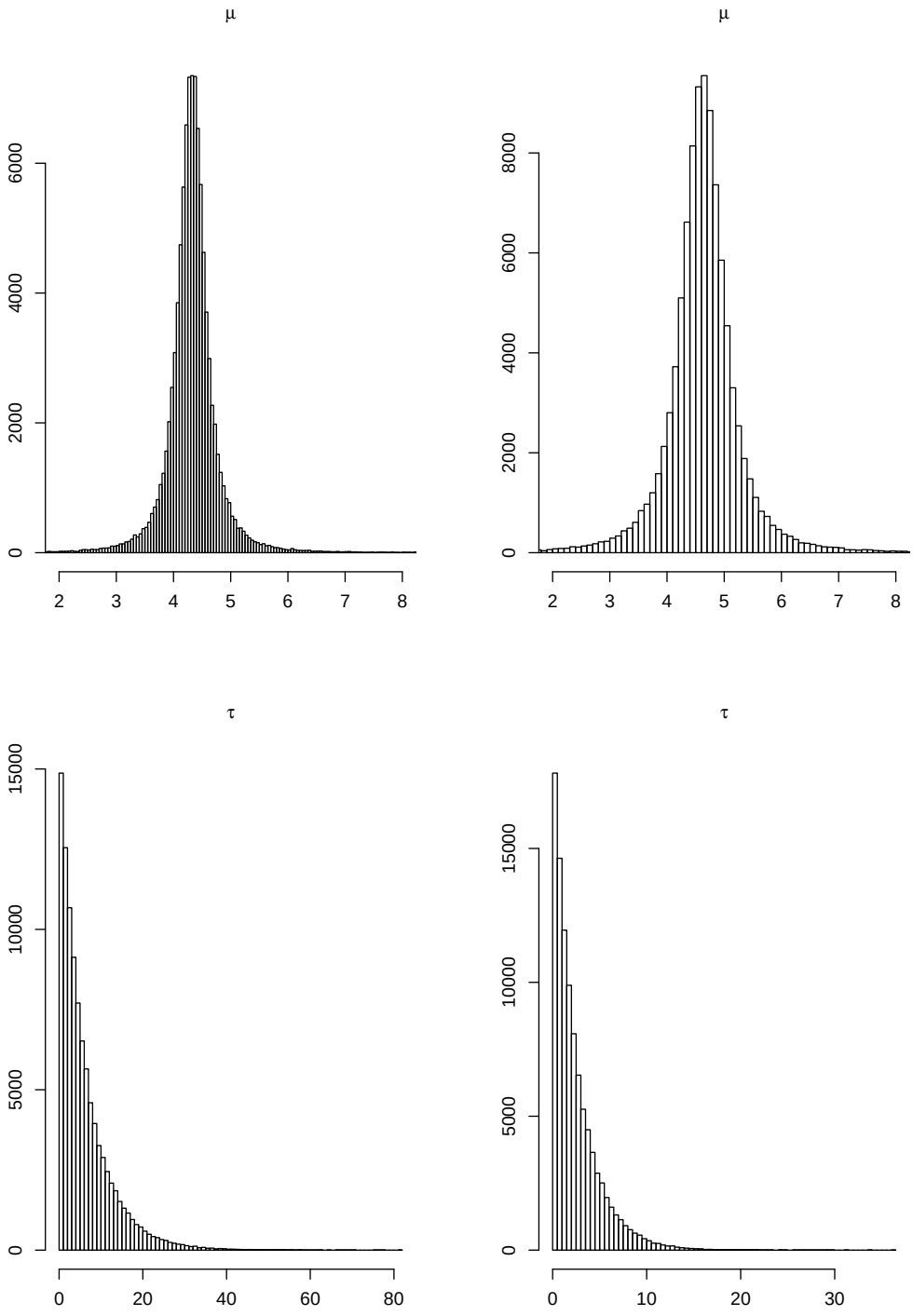


**Figure B.24:** hypermean (top) and hyperprecisions (bottom) densities of the measurement error parameters, $\sigma_N$ and $\sigma_C$, estimated via a hierarchical approach from the repeated measurements on the thee cells described in Section 4.3.

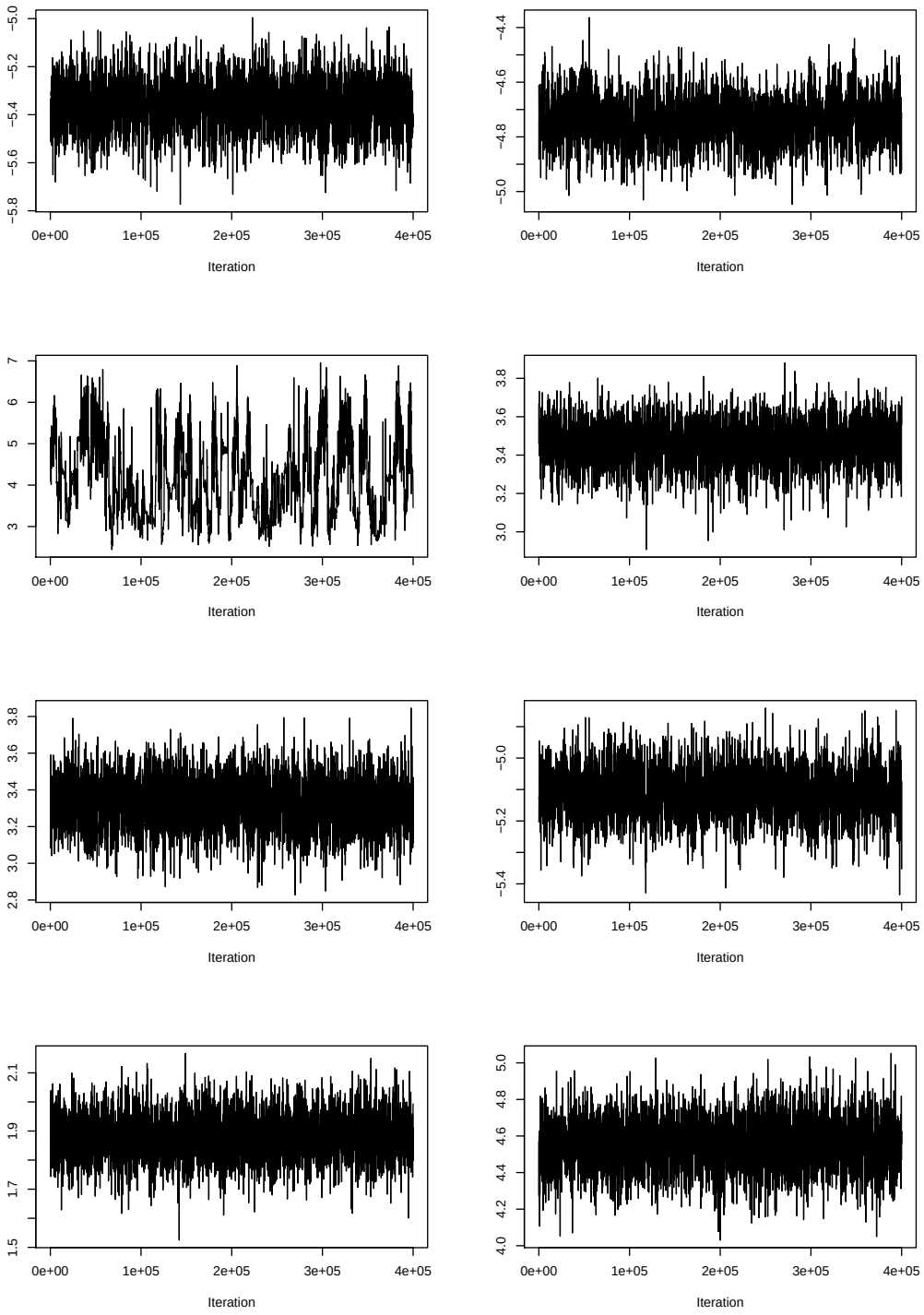


**Figure B.25:** Trace plots, excluding *burn-in*, for the posterior chains of the hypermean parameters inferred from the experimental data, under the basal condition. Figures refer, from left to right and top to bottom, to the hypermeans of $k_d^{(i)}, k_a^{(i)}, K_a^{(i)}, \mu_\tau^{(i)}, \gamma^{(i)}, \delta^{(i)}, c^{(i)}, \kappa_N^{(i)}$.

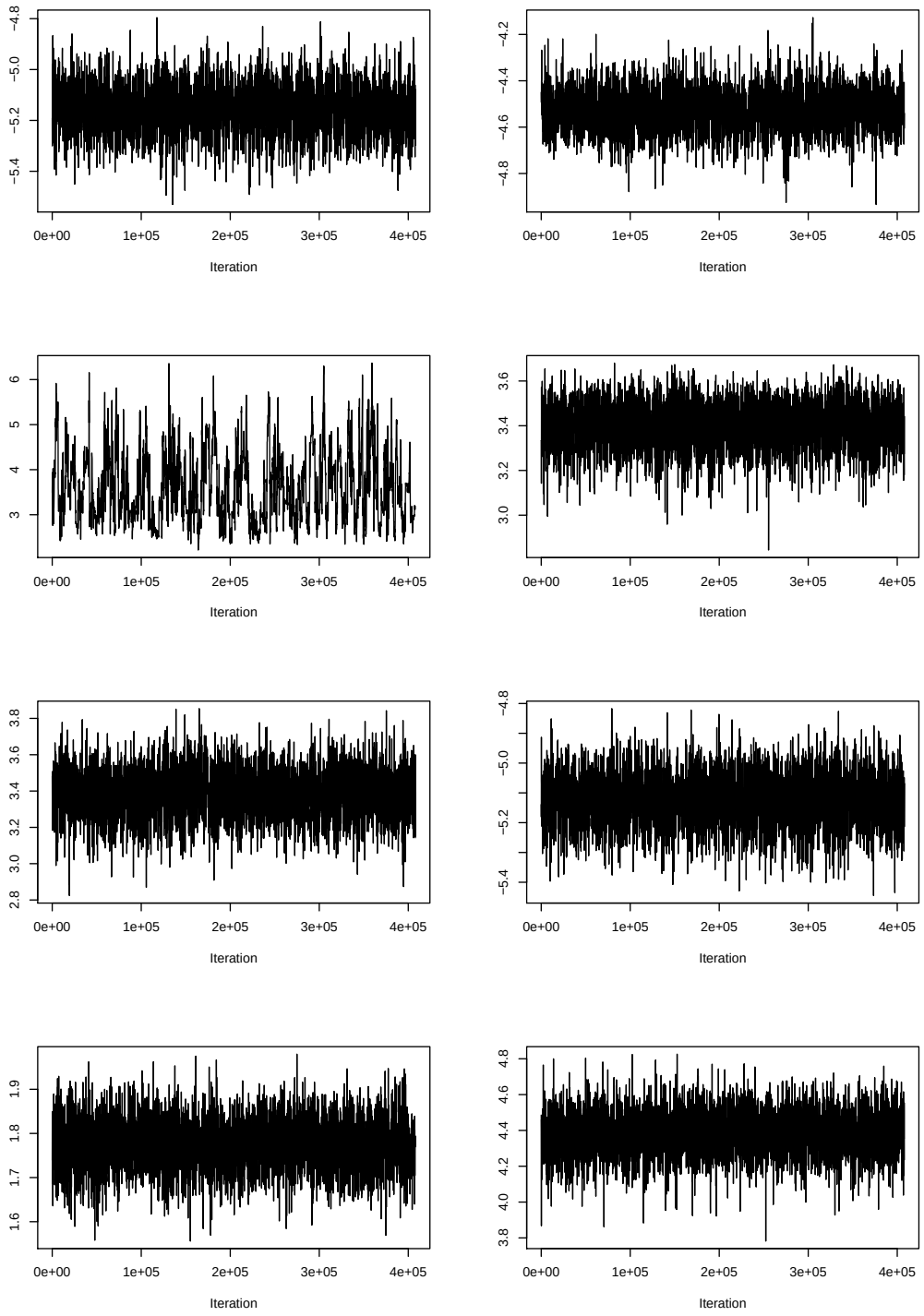


**Figure B.26:** Trace plots, excluding *burn-in*, for the posterior chains of the hypermean parameters inferred from the experimental data, under the stimulated condition. Figures refer, from left to right and top to bottom, to the hypermeans of $k_d^{(i)}, k_a^{(i)}, K_a^{(i)}, \mu_\tau^{(i)}, \gamma^{(i)}, \delta^{(i)}, c^{(i)}, \kappa_N^{(i)}$.

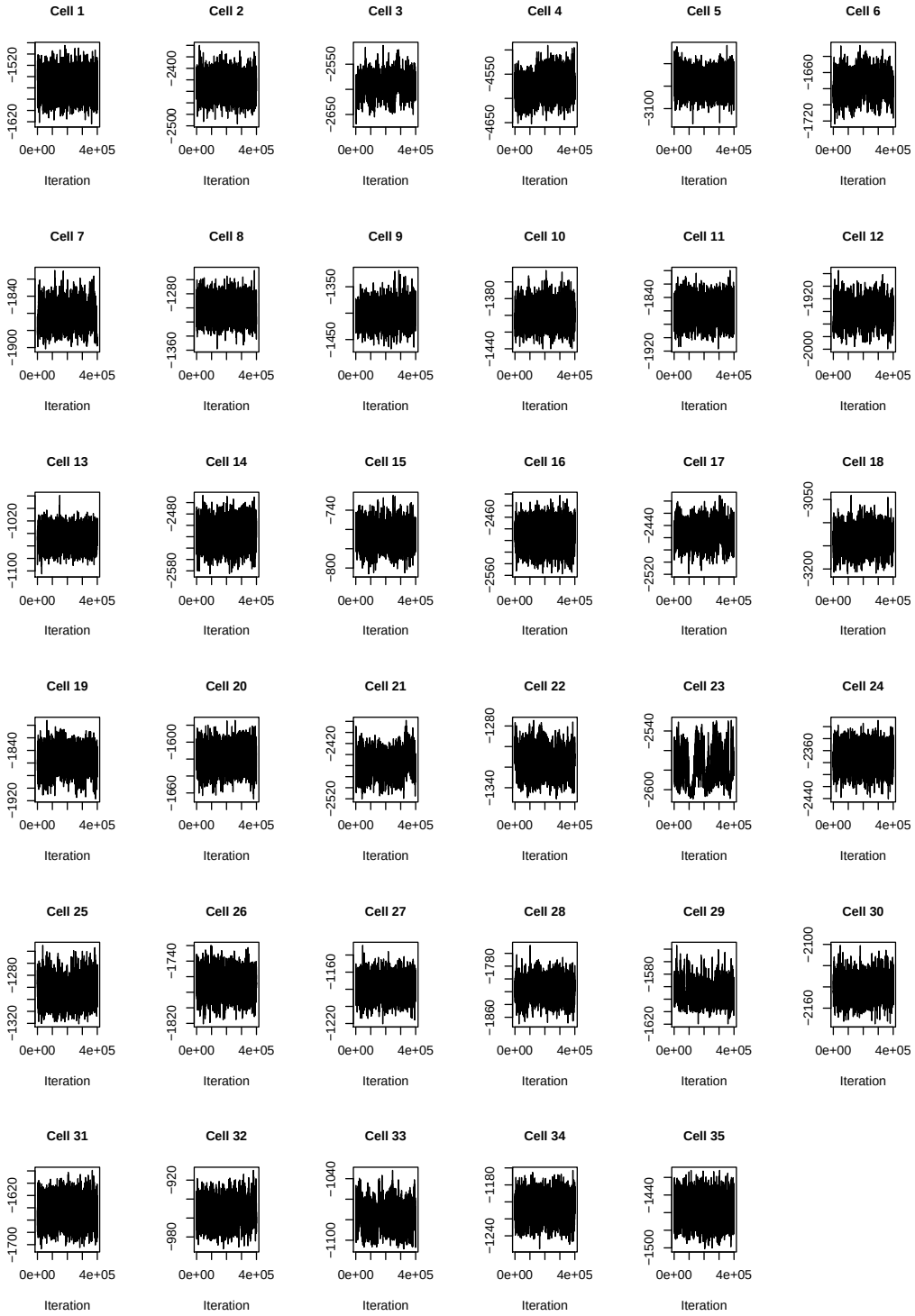


**Figure B.27:** Trace plots, excluding *burn-in*, for the posterior chains, in the log space, of the $N$ hierarchical DA likelihoods in 3.17, $log(L_{DA}(\theta^{(i)}; \tilde{x}^{(i)}))$, in each cell under the basal condition.

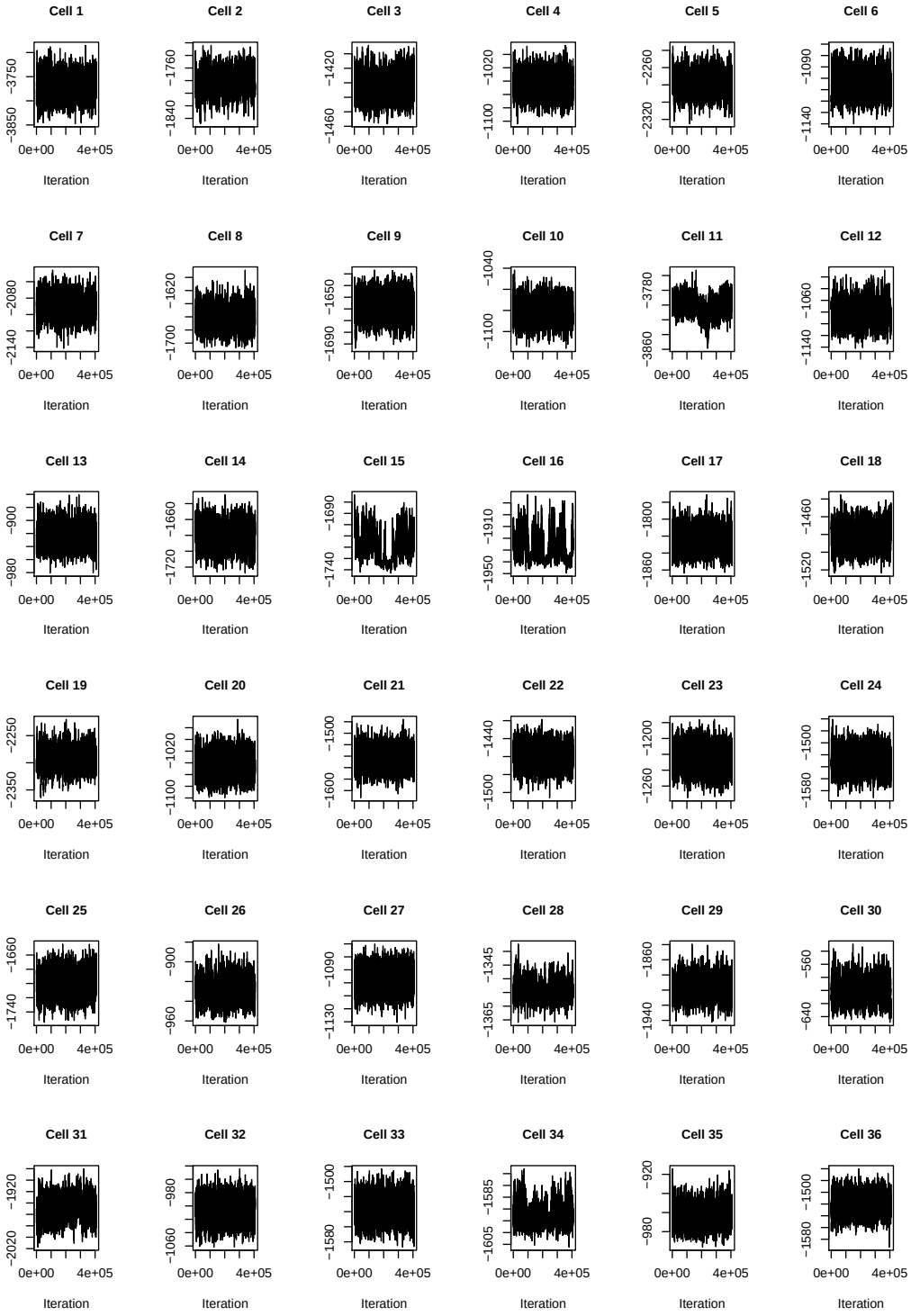


**Figure B.28:** Trace plots, excluding *burn-in*, for the posterior chains, in the log space, of the $N$ hierarchical DA likelihoods in 3.17, $log(L_{DA}(\theta^{(i)}; \tilde{x}^{(i)}))$, in each cell under the stimulated condition.

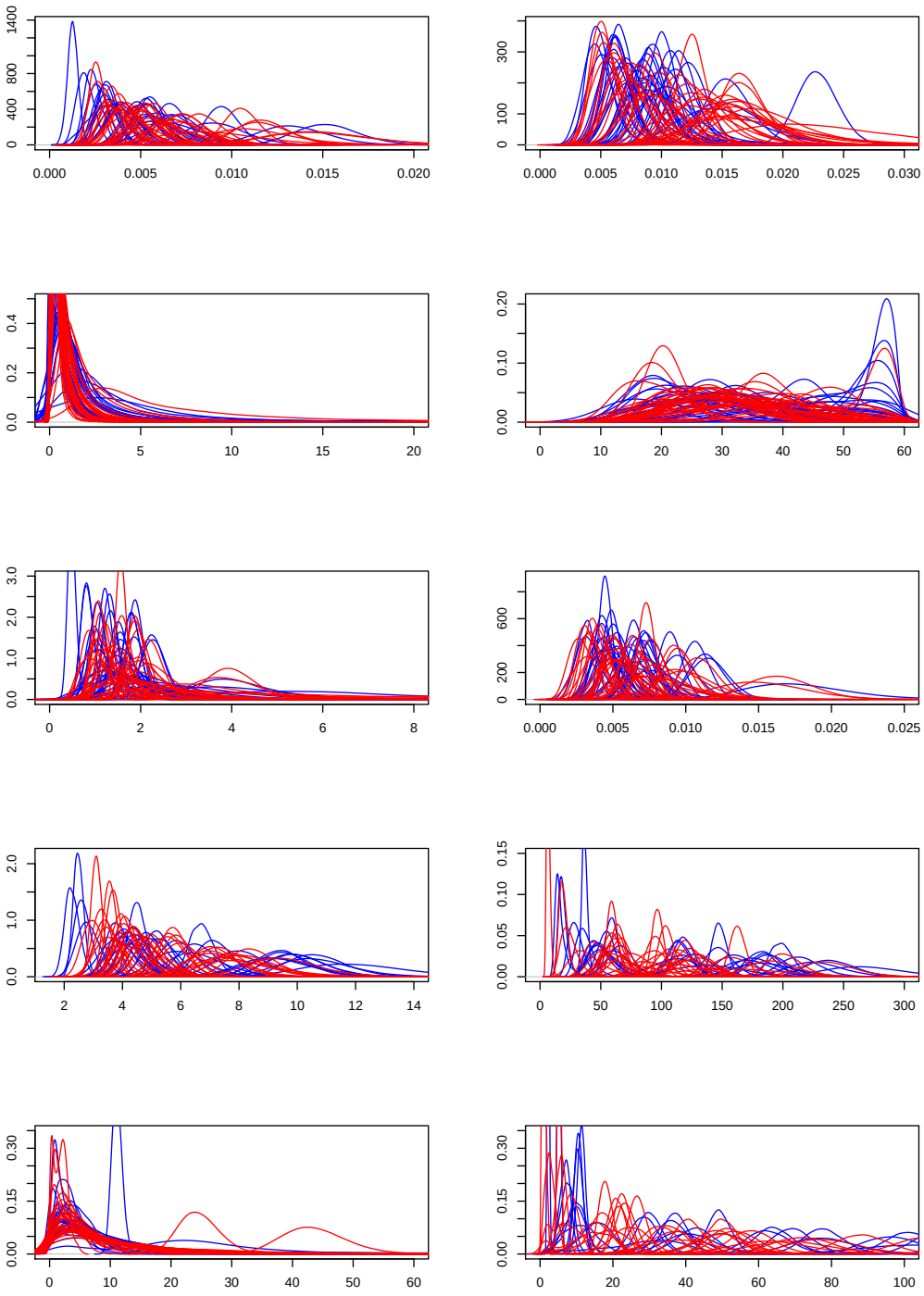


**Figure B.29:** Posterior densities for the hierarchical parameters under the basal (blue) and stimulated conditions (red). Parameters, from left to right and top to bottom, refer to: $k_d^{(i)}, k_a^{(i)}, K_a^{(i)}, \mu_\tau^{(i)}, \gamma^{(i)}, \delta^{(i)}, c^{(i)}, \kappa_N^{(i)}, \sigma_N^{(i)}, \sigma_C^{(i)}$.

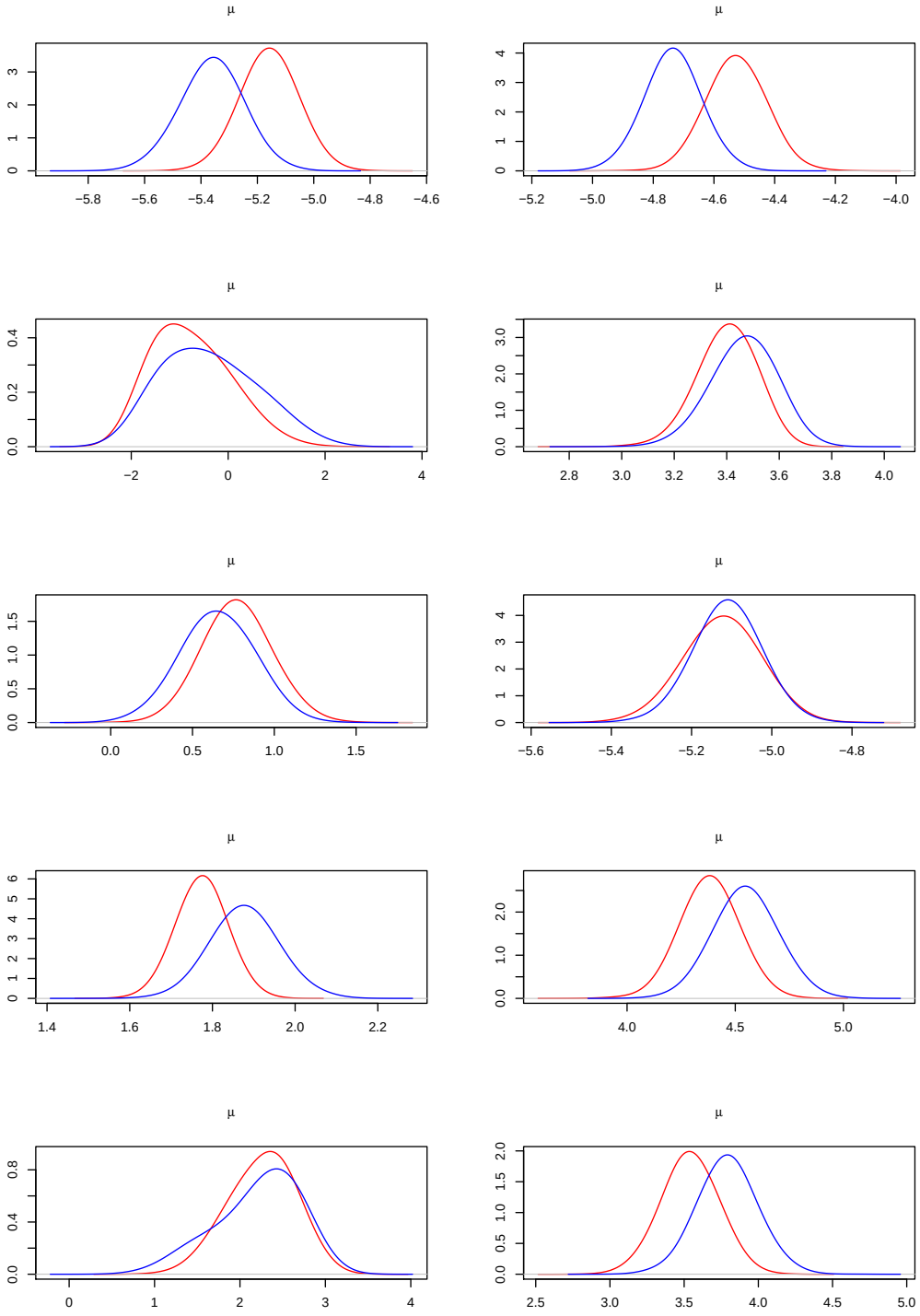


**Figure B.30:** Posterior densities for the hypermean of parameters, under the basal (blue) and stimulated conditions (red). Parameters, from left to right and top to bottom, refer to: $k_d^{(i)}, k_a^{(i)}, K_a^{(i)}, \mu_\tau^{(i)}, \gamma^{(i)}, \delta^{(i)}, c^{(i)}, \kappa_N^{(i)}, \sigma_N^{(i)}, \sigma_C^{(i)}$.

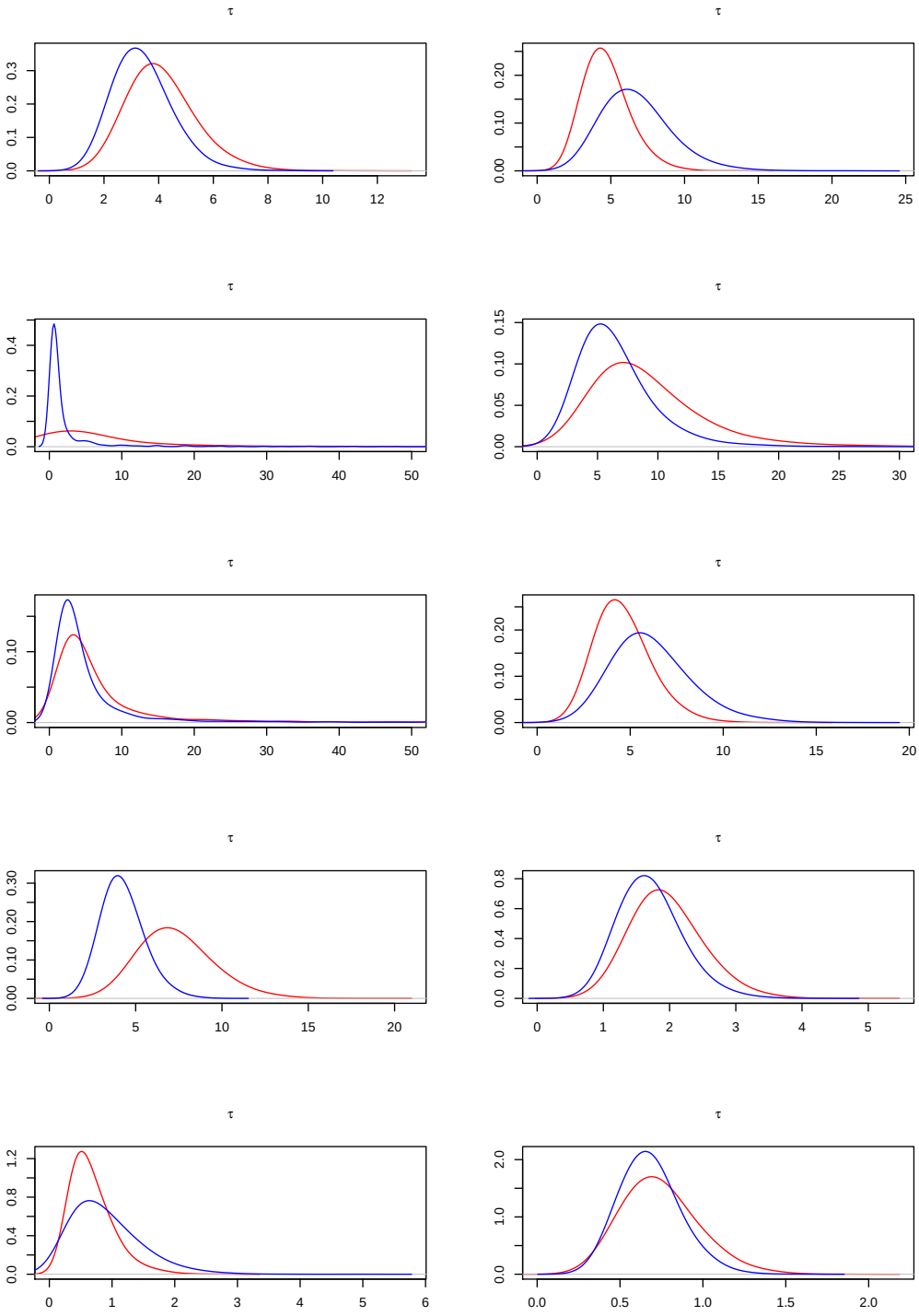


**Figure B.31:** Posterior densities for the hyperprecision parameters, under the basal (blue) and stimulated conditions (red). Parameters, from left to right and top to bottom, refer to: $k_d^{(i)}, k_a^{(i)}, K_a^{(i)}, \mu_\tau^{(i)}, \gamma^{(i)}, \delta^{(i)}, c^{(i)}, \kappa_N^{(i)}, \sigma_N^{(i)}, \sigma_C^{(i)}$.

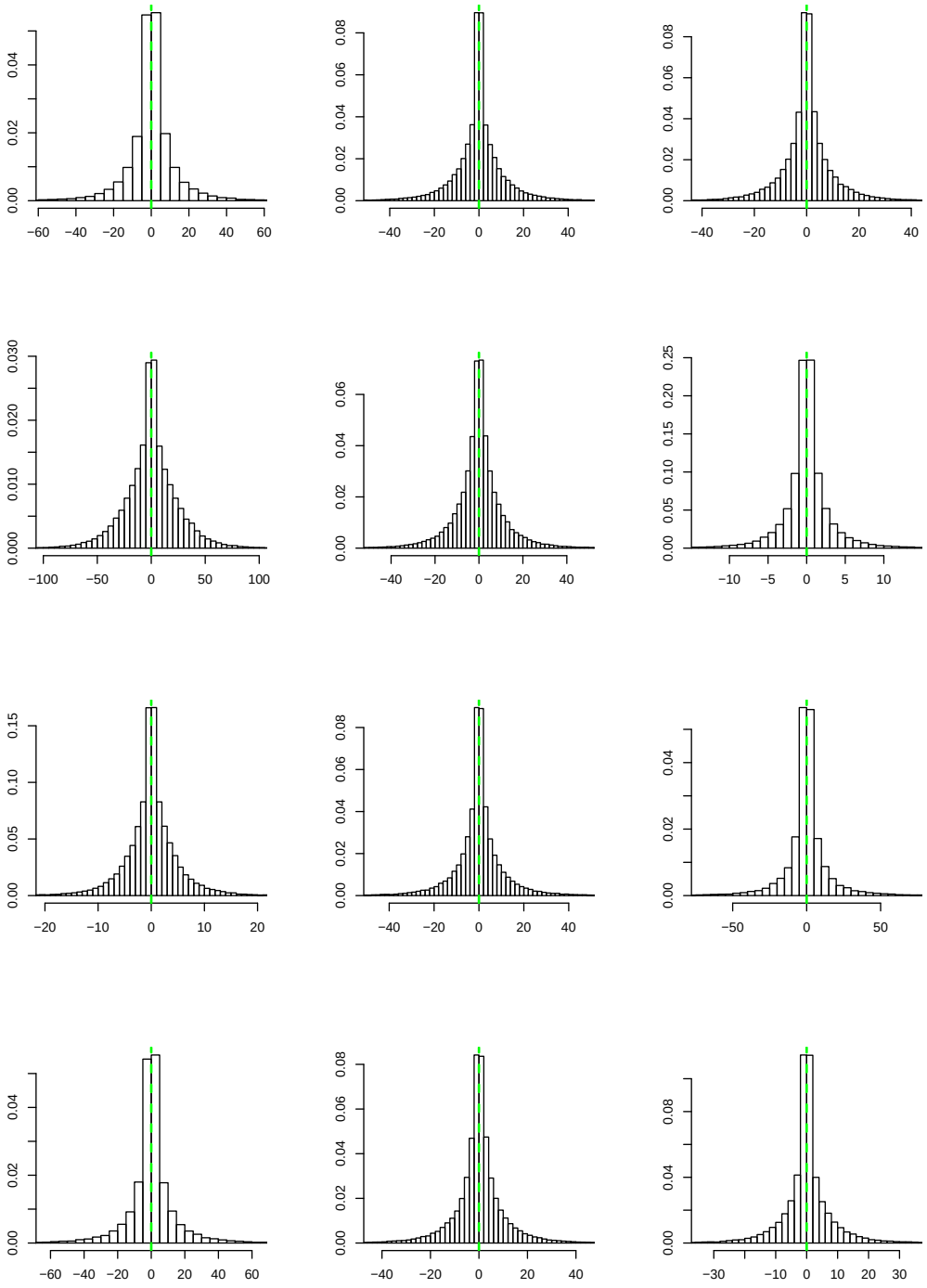


**Figure B.32:** Histograms of the nuclear residuals over the entire process, for cells from 1 to 12 under the basal condition.

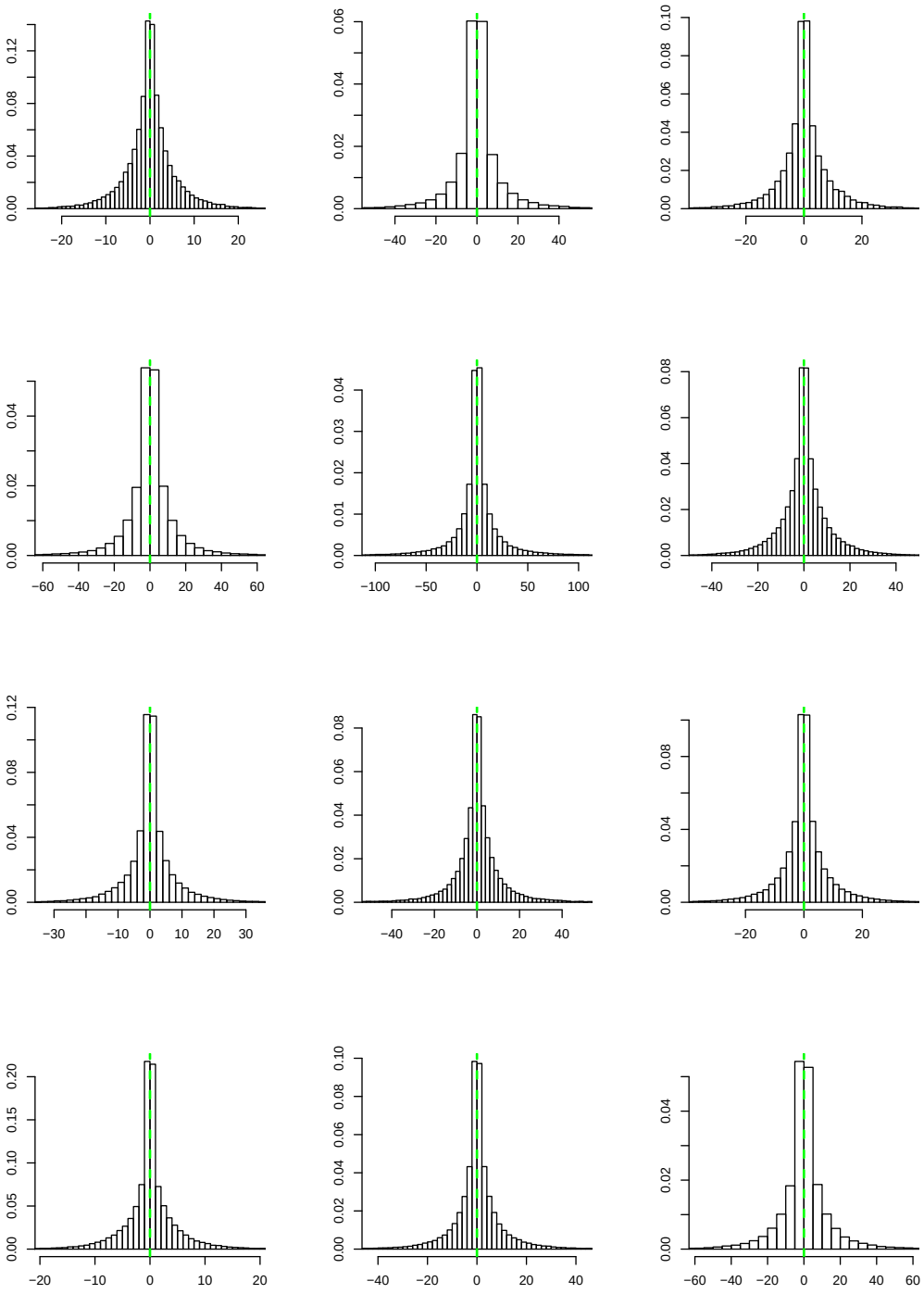


**Figure B.33:** Histograms of the nuclear residuals over the entire process, for cells from 13 to 24 under the basal condition.

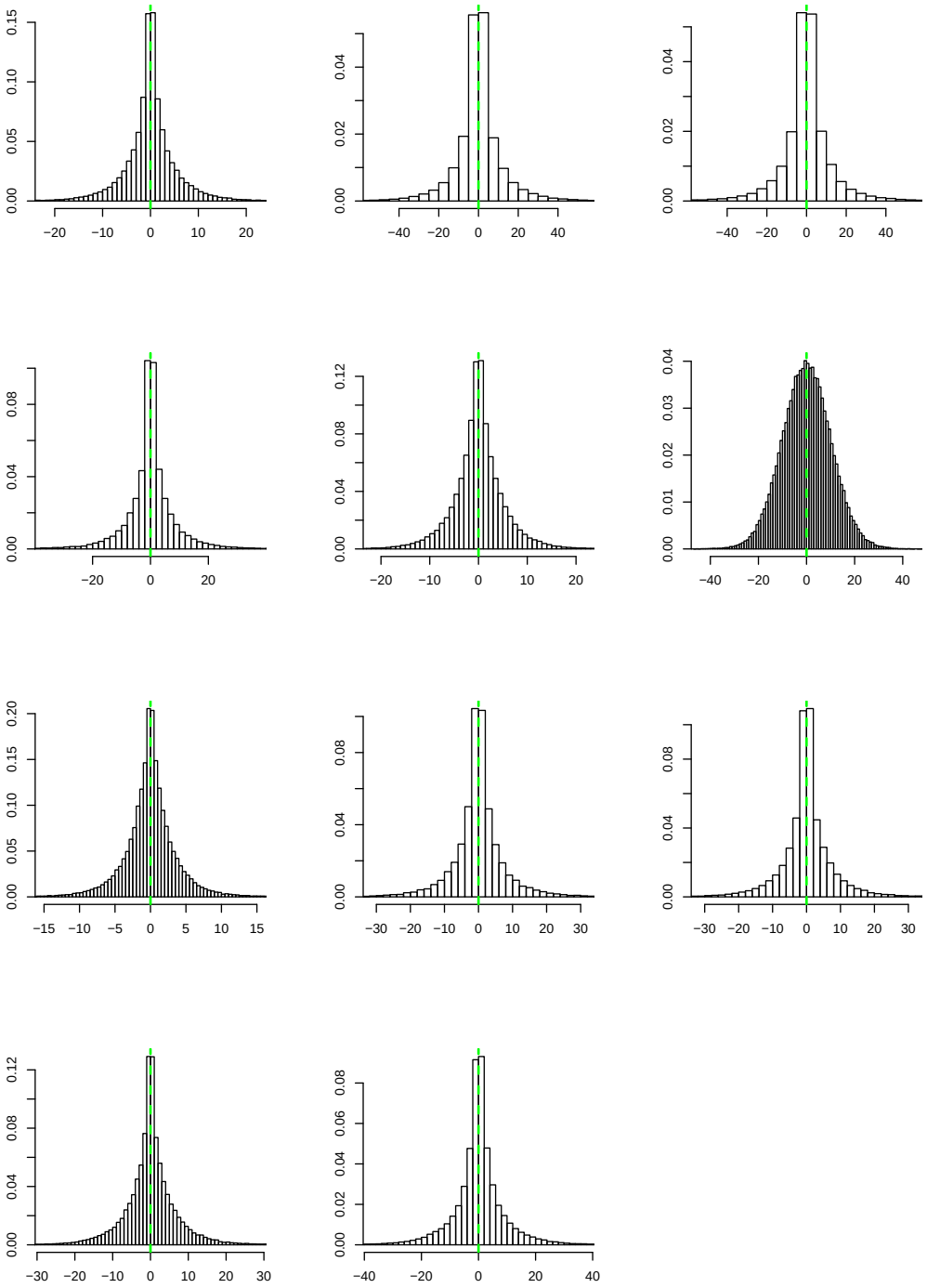

**Figure B.34:** Histograms of the nuclear residuals over the entire process, for cells from 25 to 35 under the basal condition.

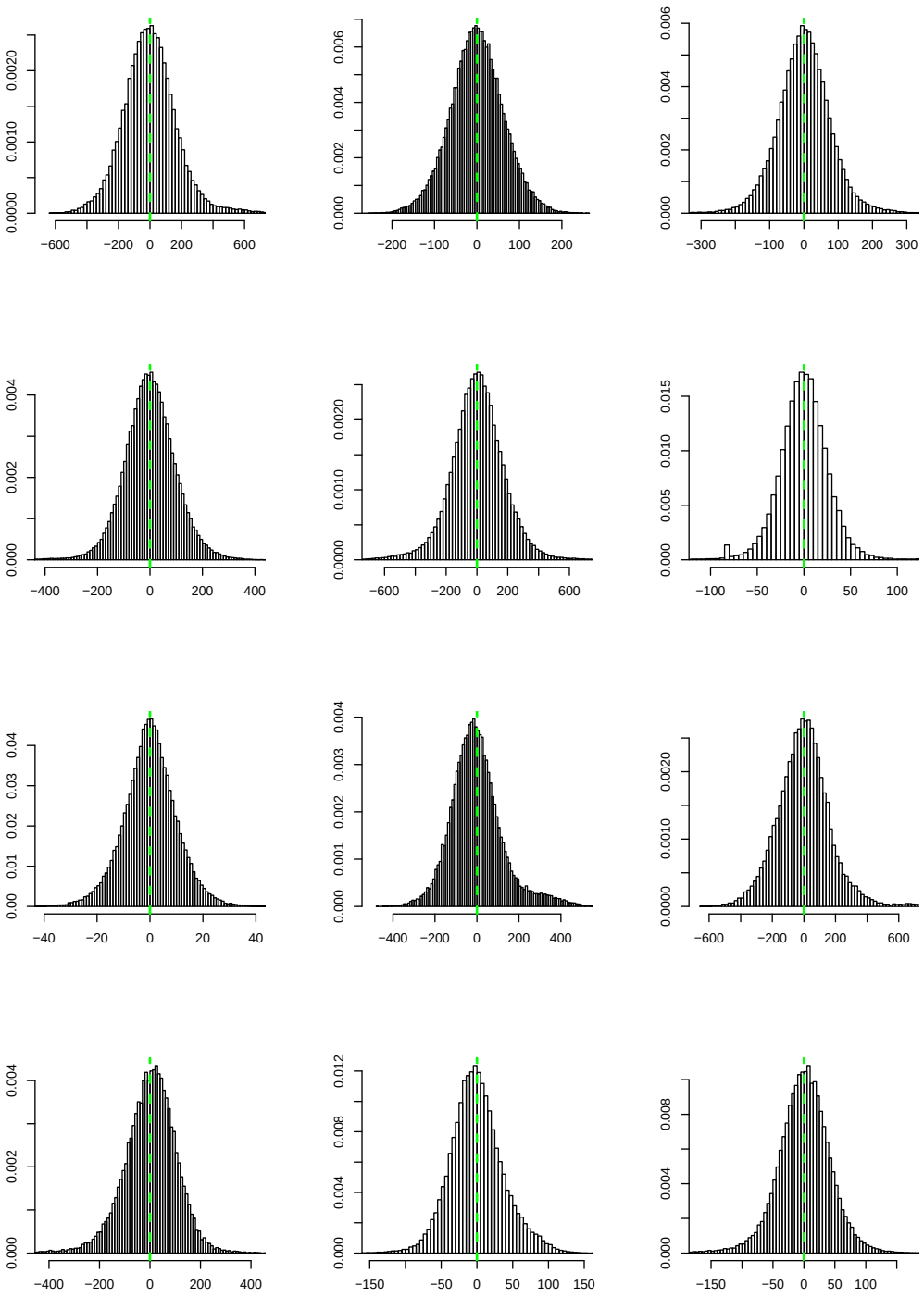


**Figure B.35:** Histograms of the cytoplasmic residuals over the entire process, for cells from 1 to 12 under the basal condition.

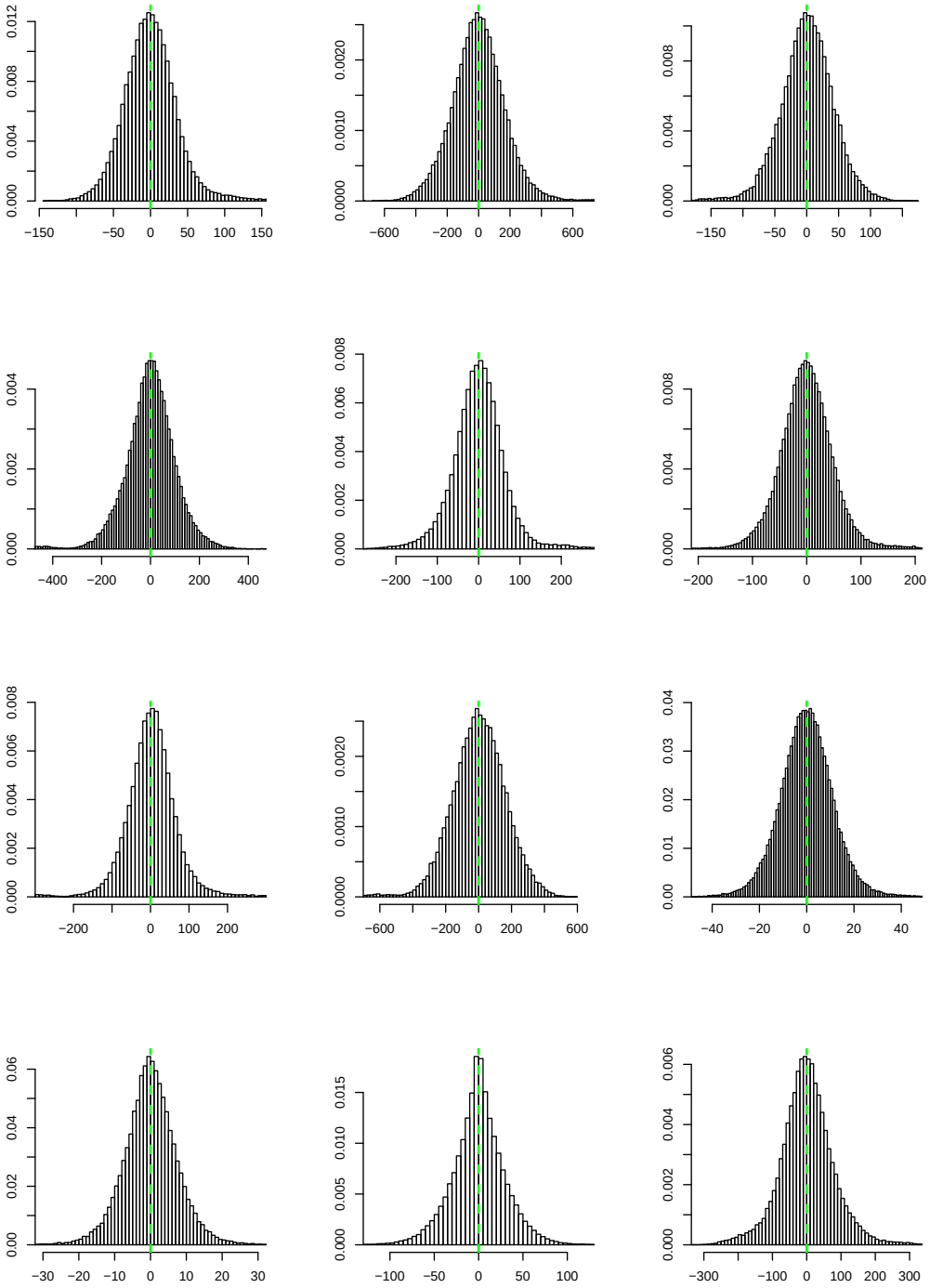


**Figure B.36:** Histograms of the cytoplasmic residuals over the entire process, for cells from 13 to 24 under the basal condition.

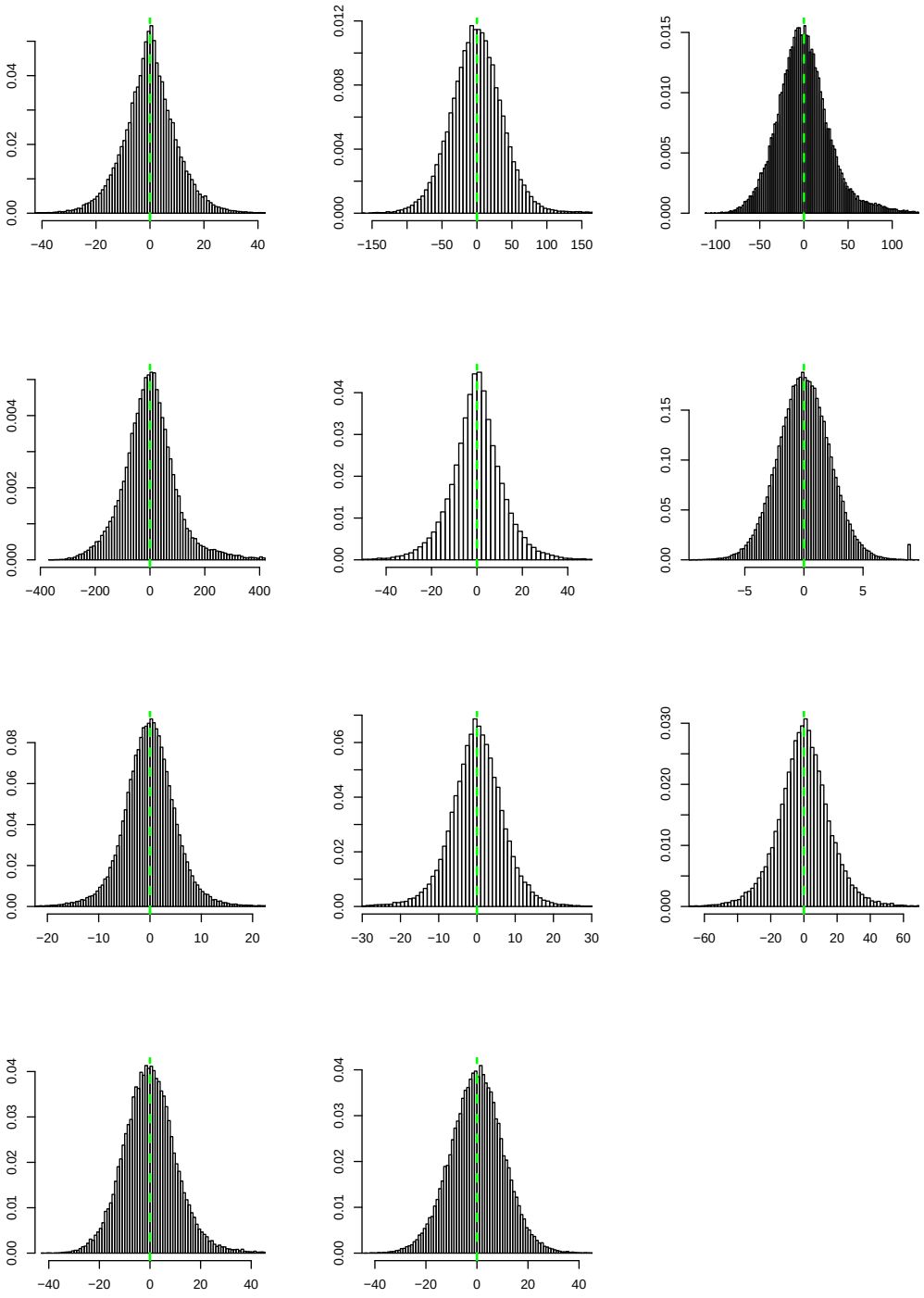


**Figure B.37:** Histograms of the cytoplasmic residuals over the entire process, for cells from 25 to 35 under the basal condition.

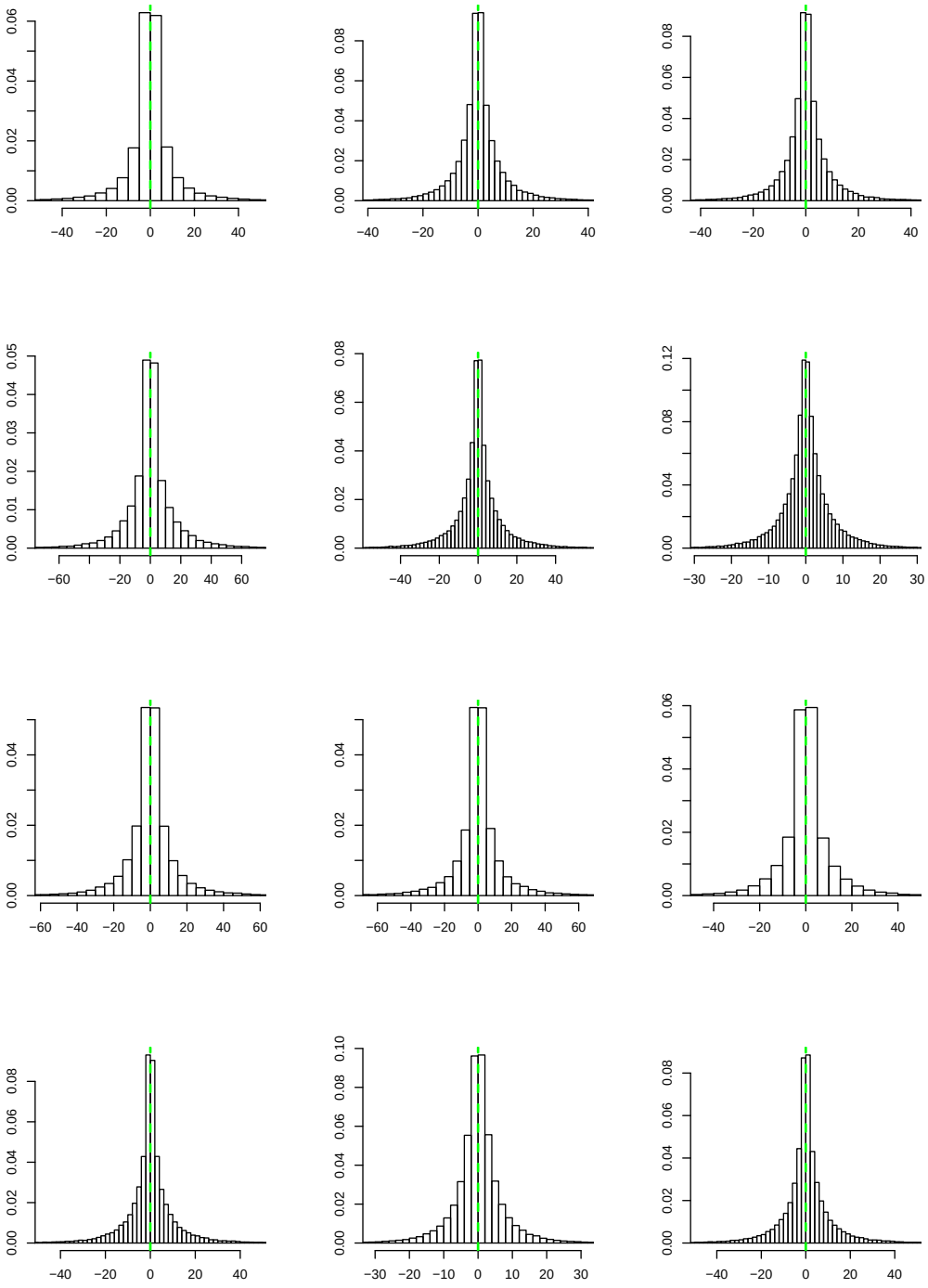

**Figure B.38:** Histograms of the nuclear residuals over the entire process, for cells from 1 to 12 under the stimulated condition.

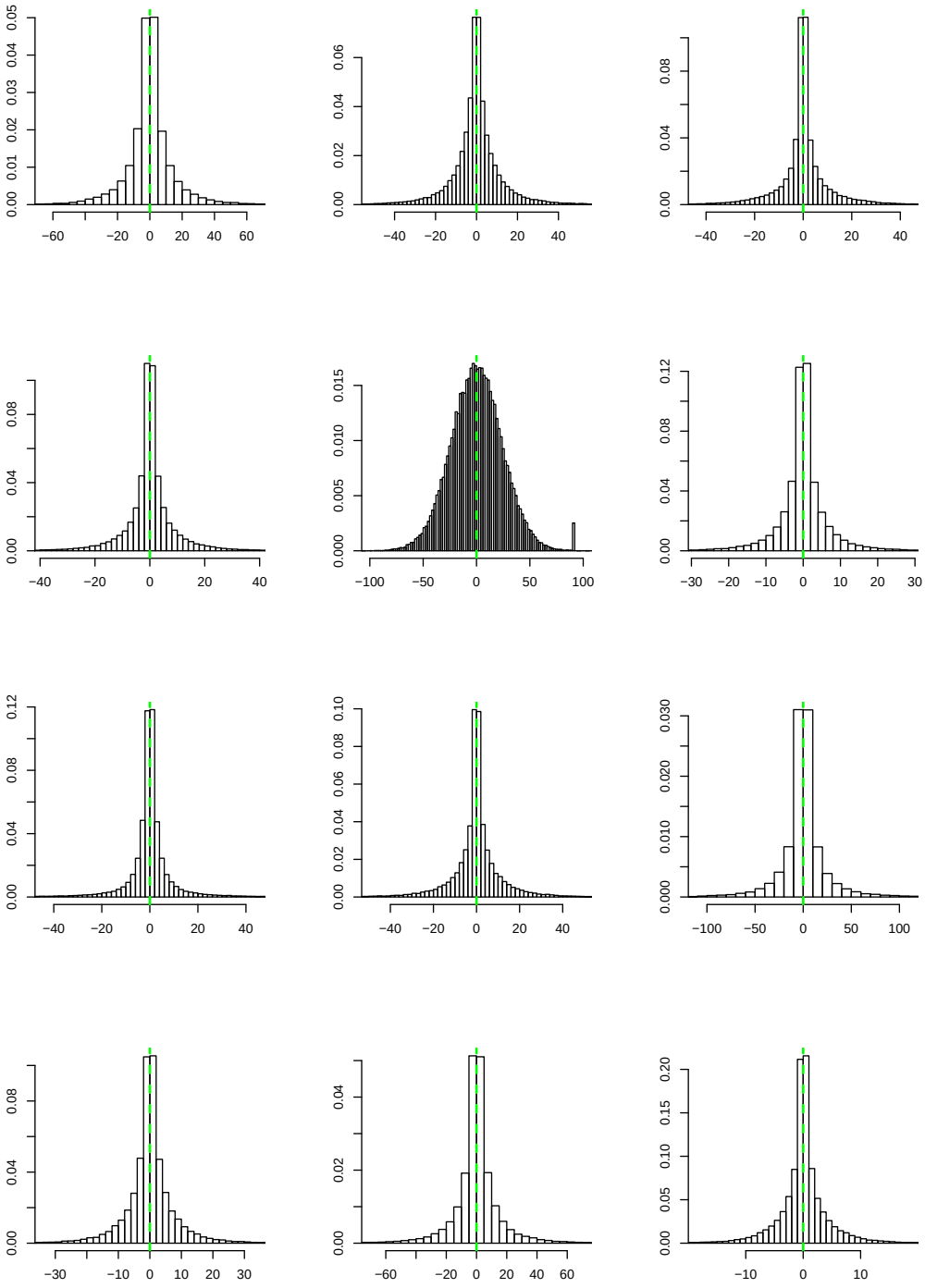


**Figure B.39:** Histograms of the nuclear residuals over the entire process, for cells from 13 to 24 under the stimulated condition.

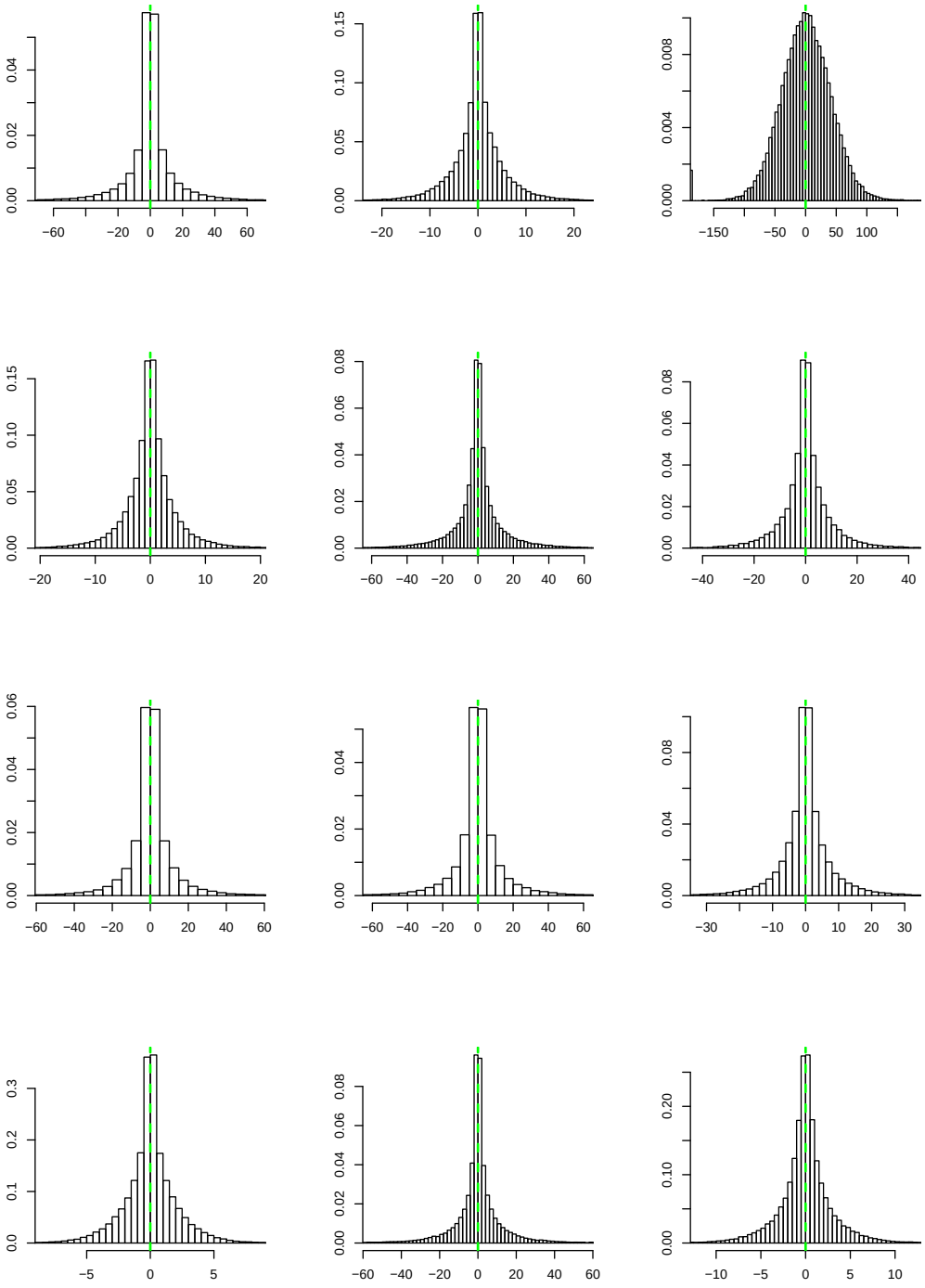


**Figure B.40:** Histograms of the nuclear residuals over the entire process, for cells from 25 to 35 under the stimulated condition.

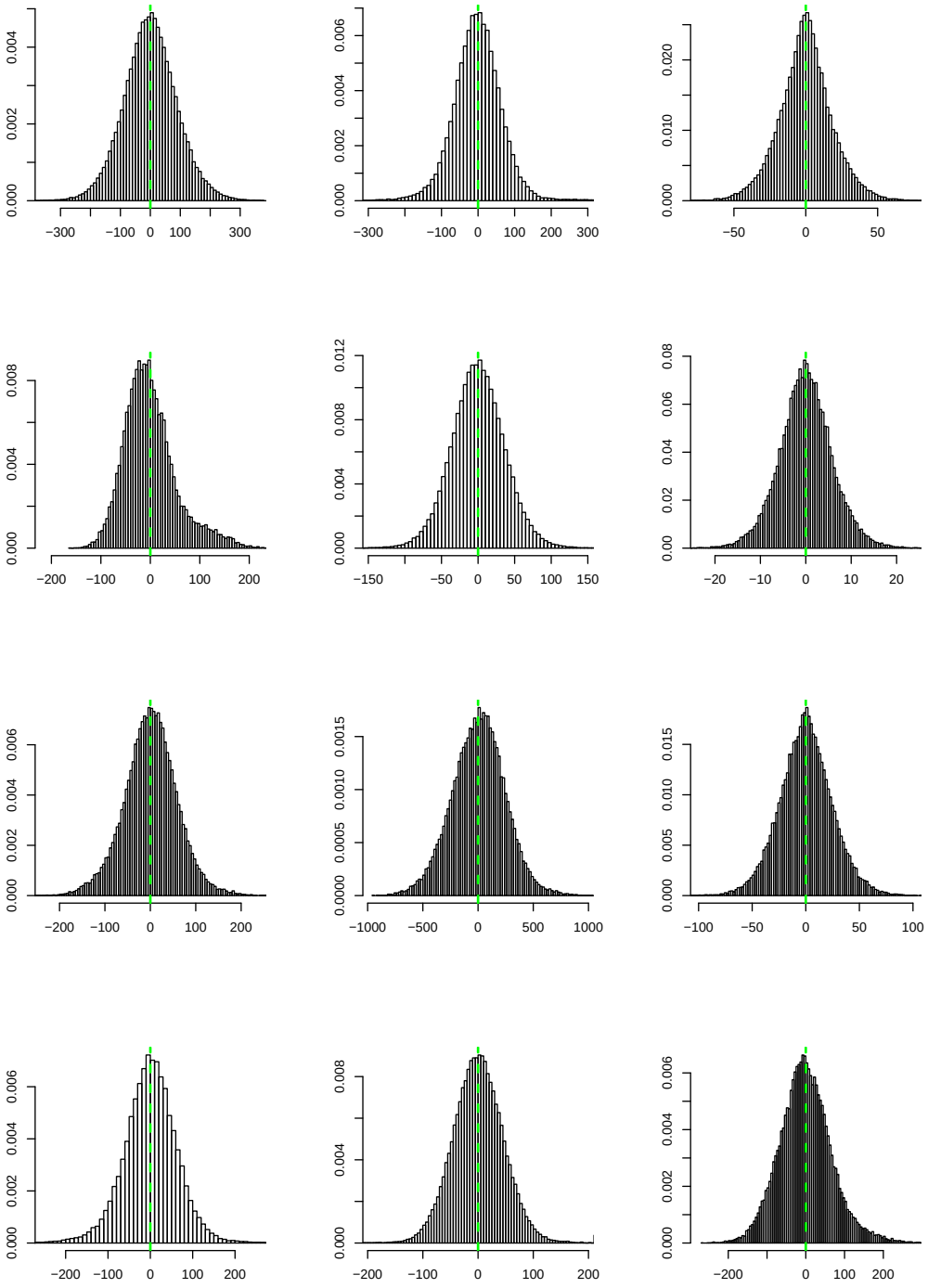


**Figure B.41:** Histograms of the cytoplasmic residuals over the entire process, for cells from 1 to 12 under the stimulated condition.

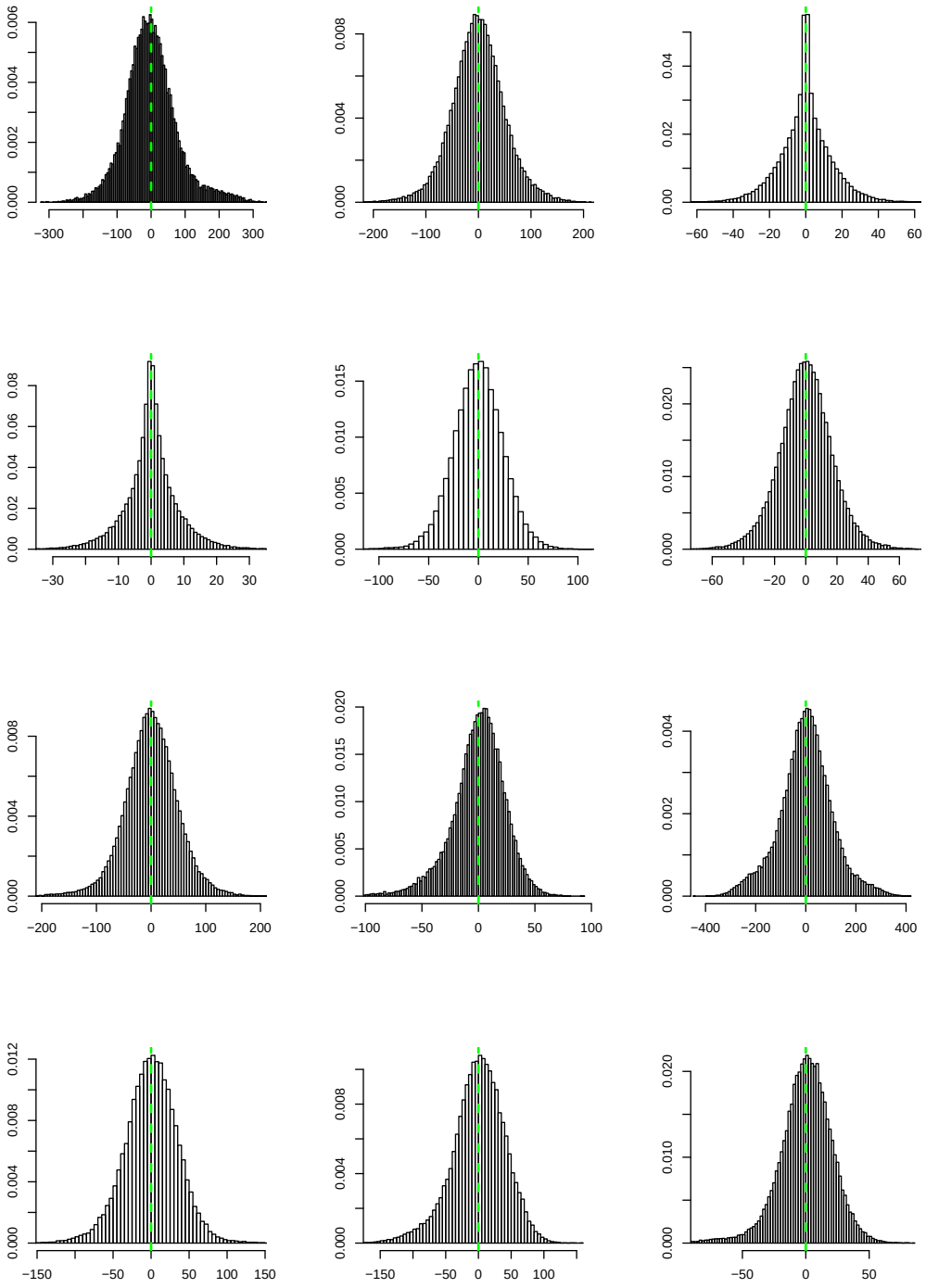


**Figure B.42:** Histograms of the cytoplasmic residuals over the entire process, for cells from 13 to 24 under the stimulated condition.

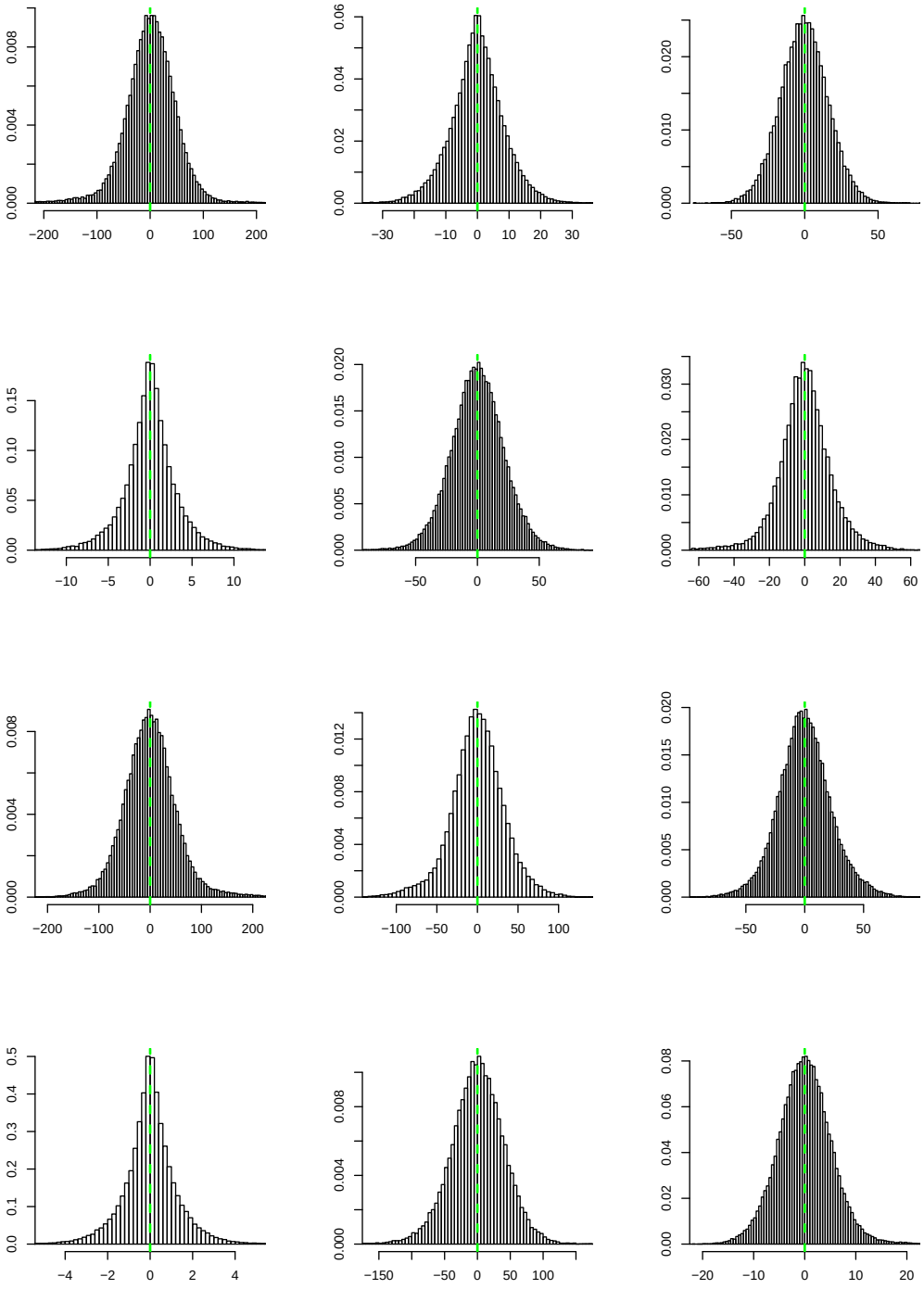


**Figure B.43:** Histograms of the cytoplasmic residuals over the entire process, for cells from 25 to 36 under the stimulated condition.

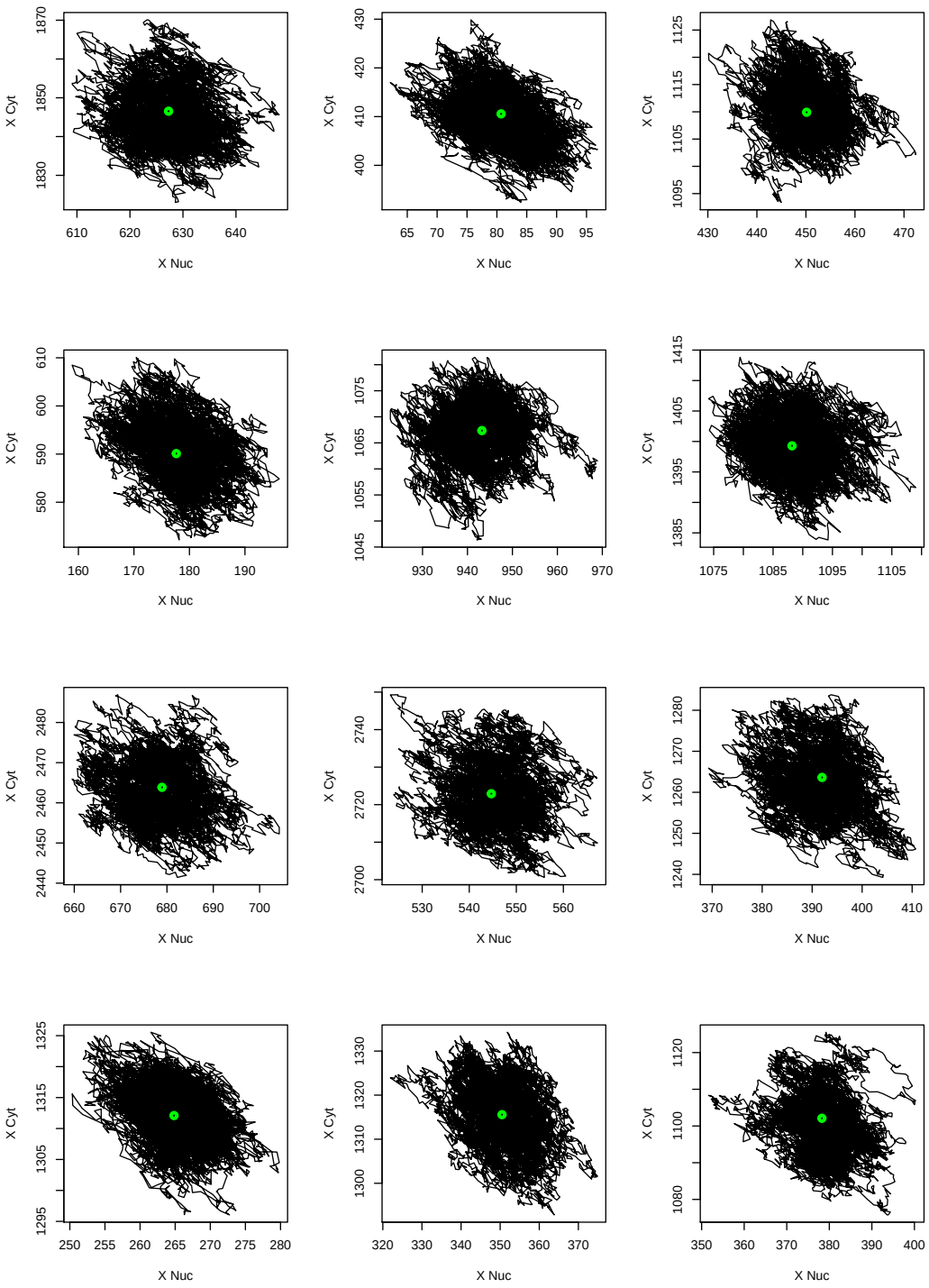


**Figure B.44:** Graphical representation, with nuclear and cytoplasmic processes on the two axes, of Euler simulations perturbed with noise; each simulation moves around the respective DDE solution, represented by the green dot. Images refer to simulations from the posterior modes of parameters obtained from cells 1-12 under the basal condition.

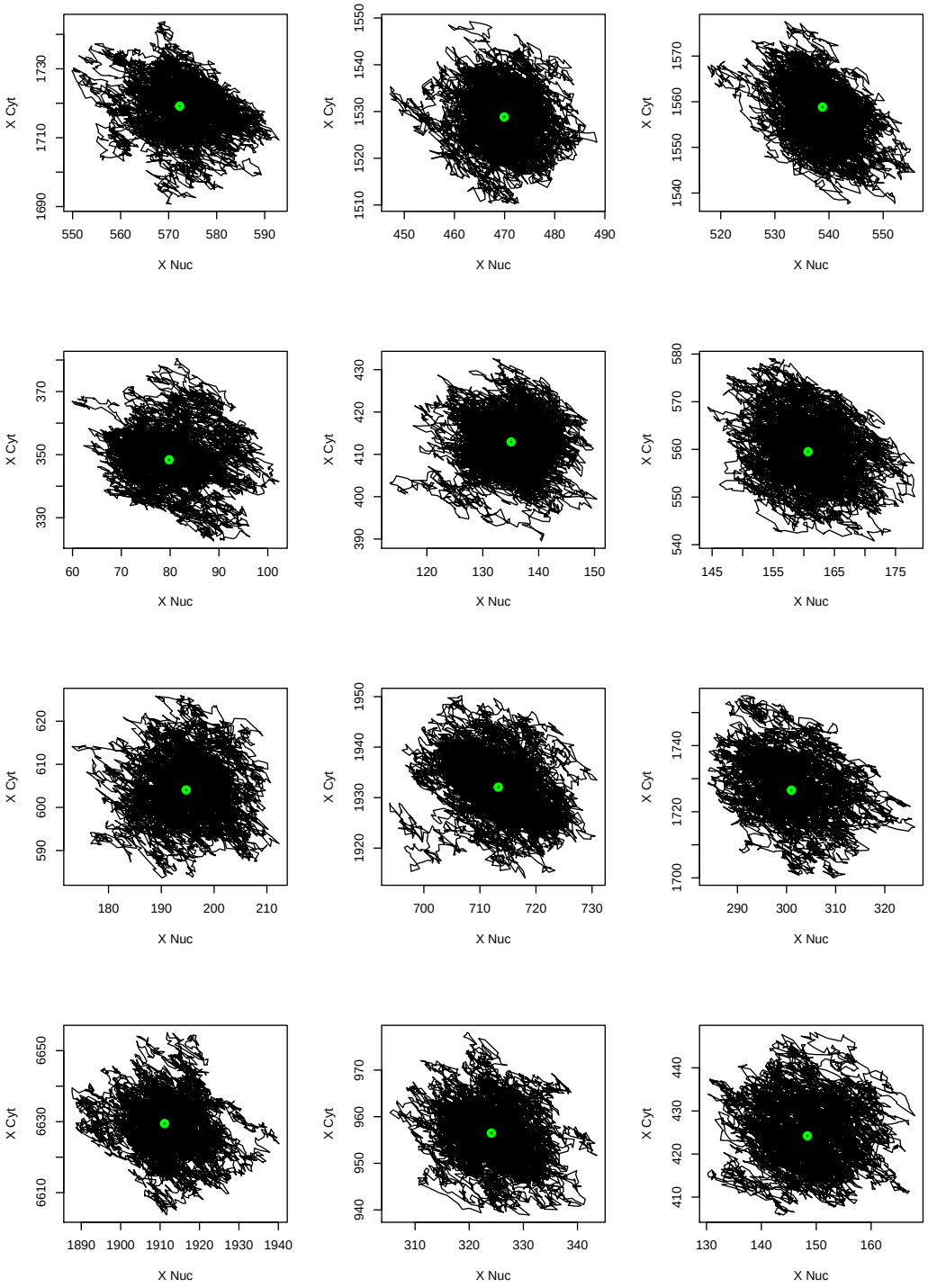


**Figure B.45:** Graphical representation, with nuclear and cytoplasmic processes on the two axes, of Euler simulations perturbed with noise; each simulation moves around the respective DDE solution, represented by the green dot. Images refer to simulations from the posterior modes of parameters obtained from cells 13-24 under the basal condition.

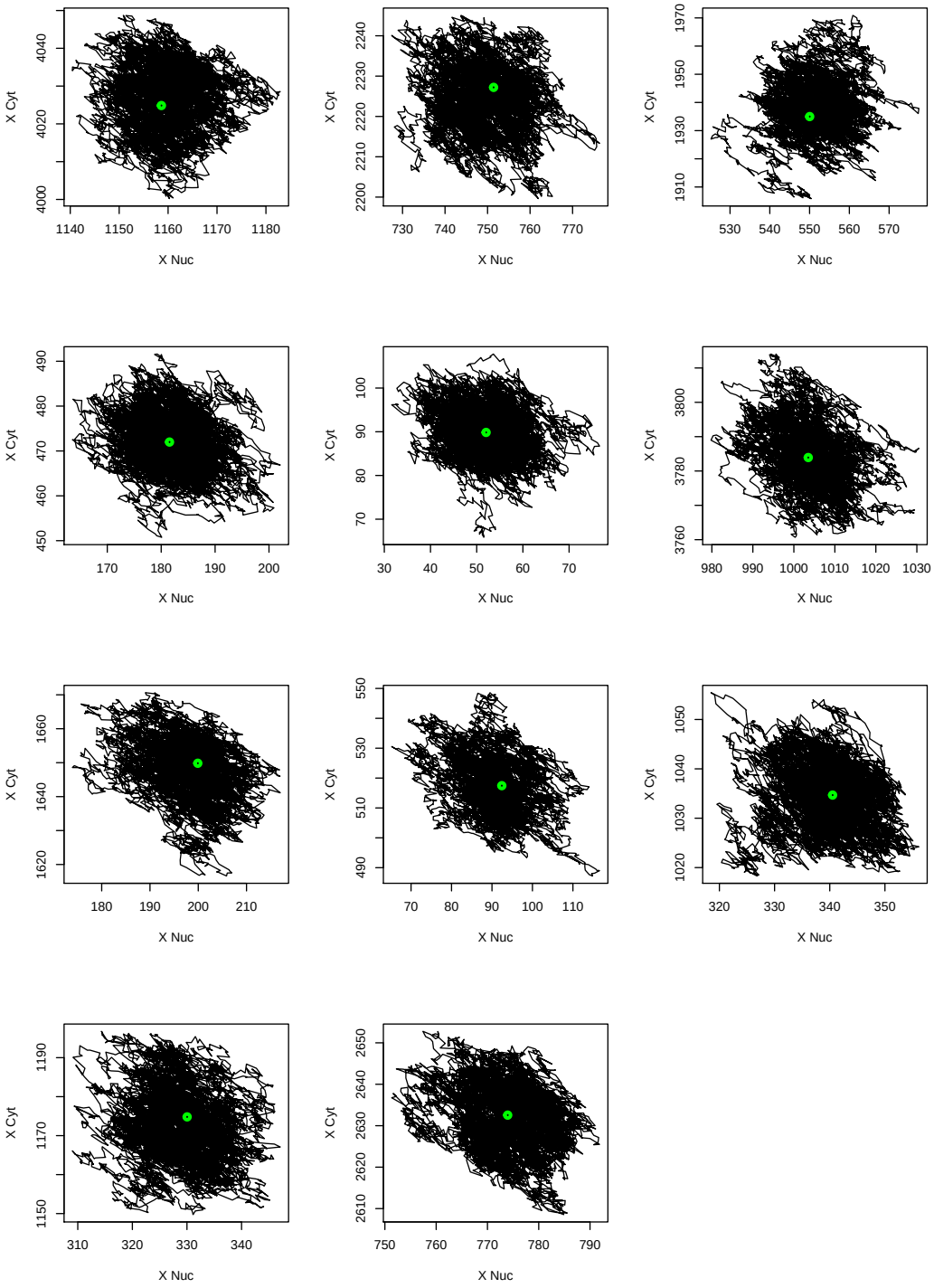


**Figure B.46:** Graphical representation, with nuclear and cytoplasmic processes on the two axes, of Euler simulations perturbed with noise; each simulation moves around the respective DDE solution, represented by the green dot. Images refer to simulations from the posterior modes of parameters obtained from cells 25-35 under the basal condition.

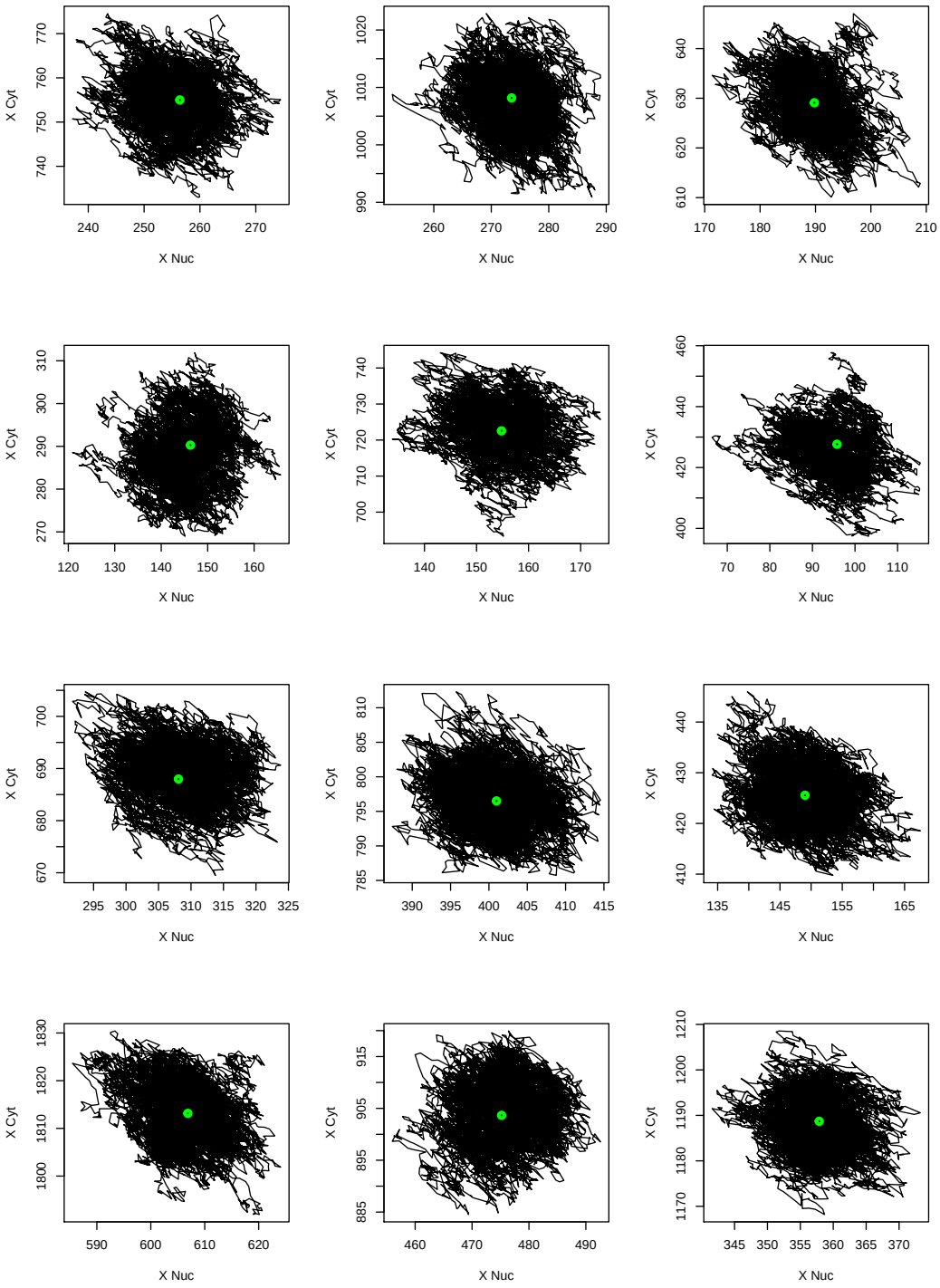


**Figure B.47:** Graphical representation, with nuclear and cytoplasmic processes on the two axes, of Euler simulations perturbed with noise; each simulation moves around the respective DDE solution, represented by the green dot. Images refer to simulations from the posterior modes of parameters obtained from cells 1-12 under the stimulated condition.

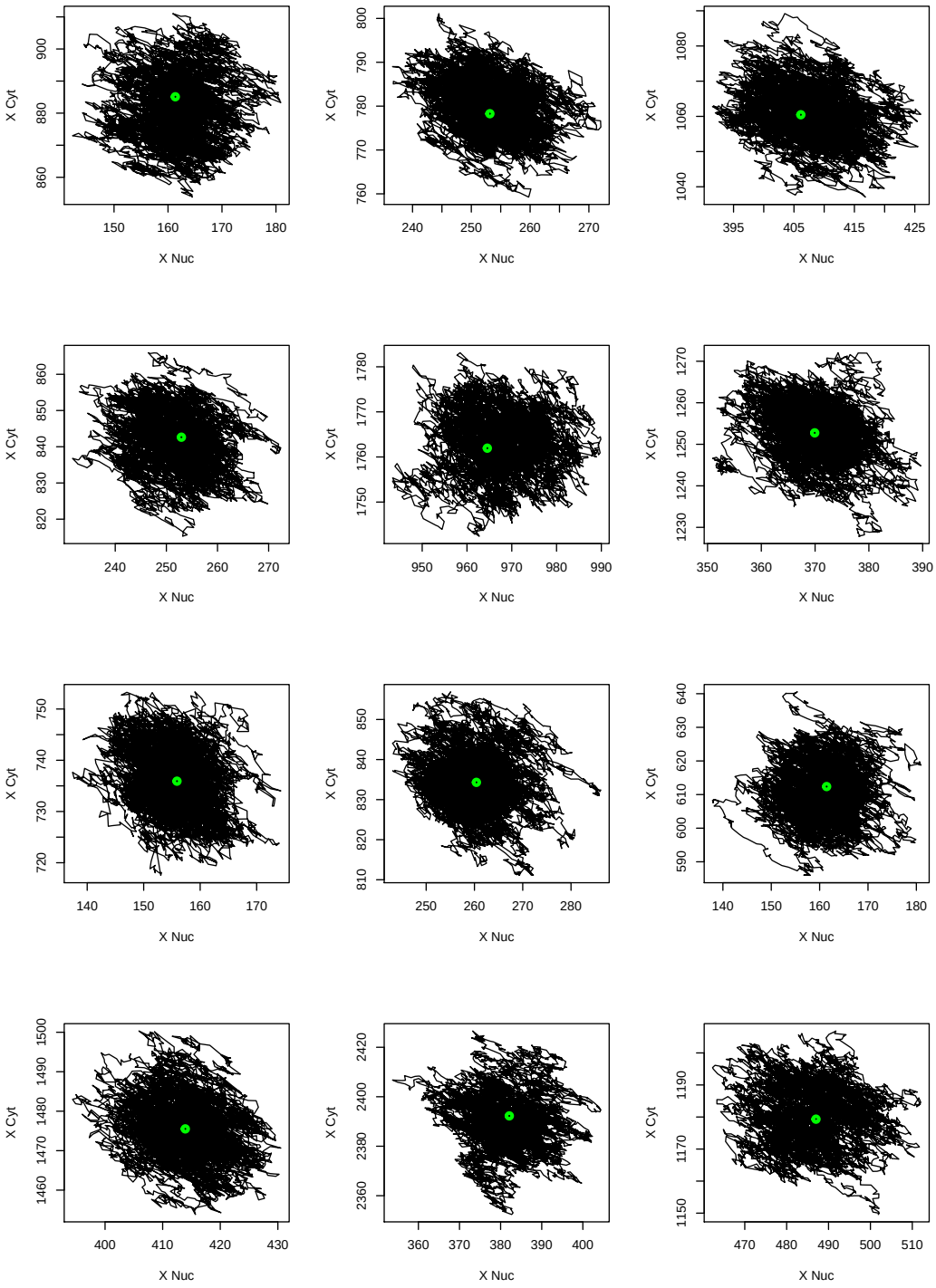


**Figure B.48:** Graphical representation, with nuclear and cytoplasmic processes on the two axes, of Euler simulations perturbed with noise; each simulation moves around the respective DDE solution, represented by the green dot. Images refer to simulations from the posterior modes of parameters obtained from cells 13-24 under the stimulated condition.

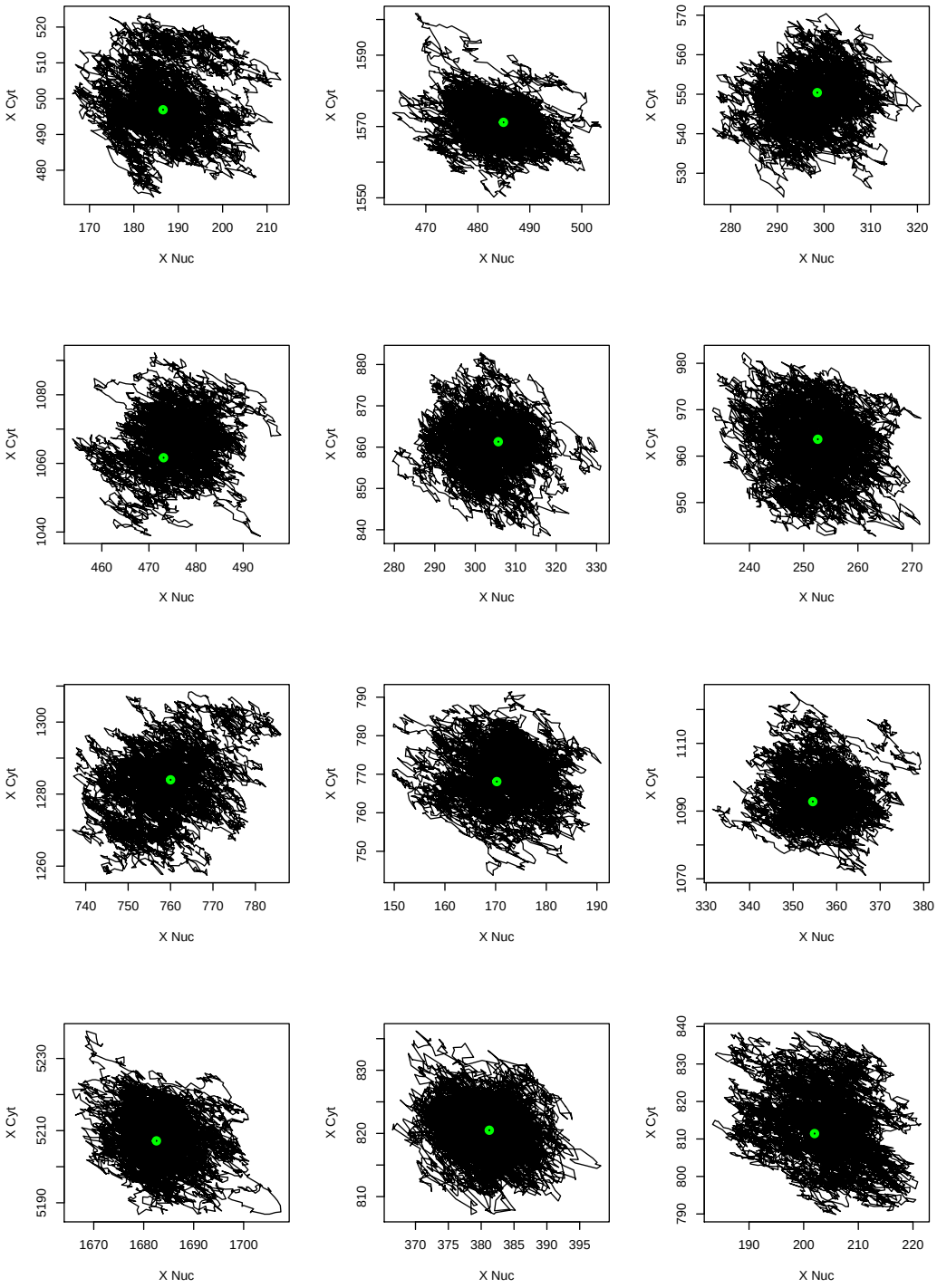


**Figure B.49:** Graphical representation, with nuclear and cytoplasmic processes on the two axes, of Euler simulations perturbed with noise; each simulation moves around the respective DDE solution, represented by the green dot. Images refer to simulations from the posterior modes of parameters obtained from cells 25-36 under the stimulated condition.

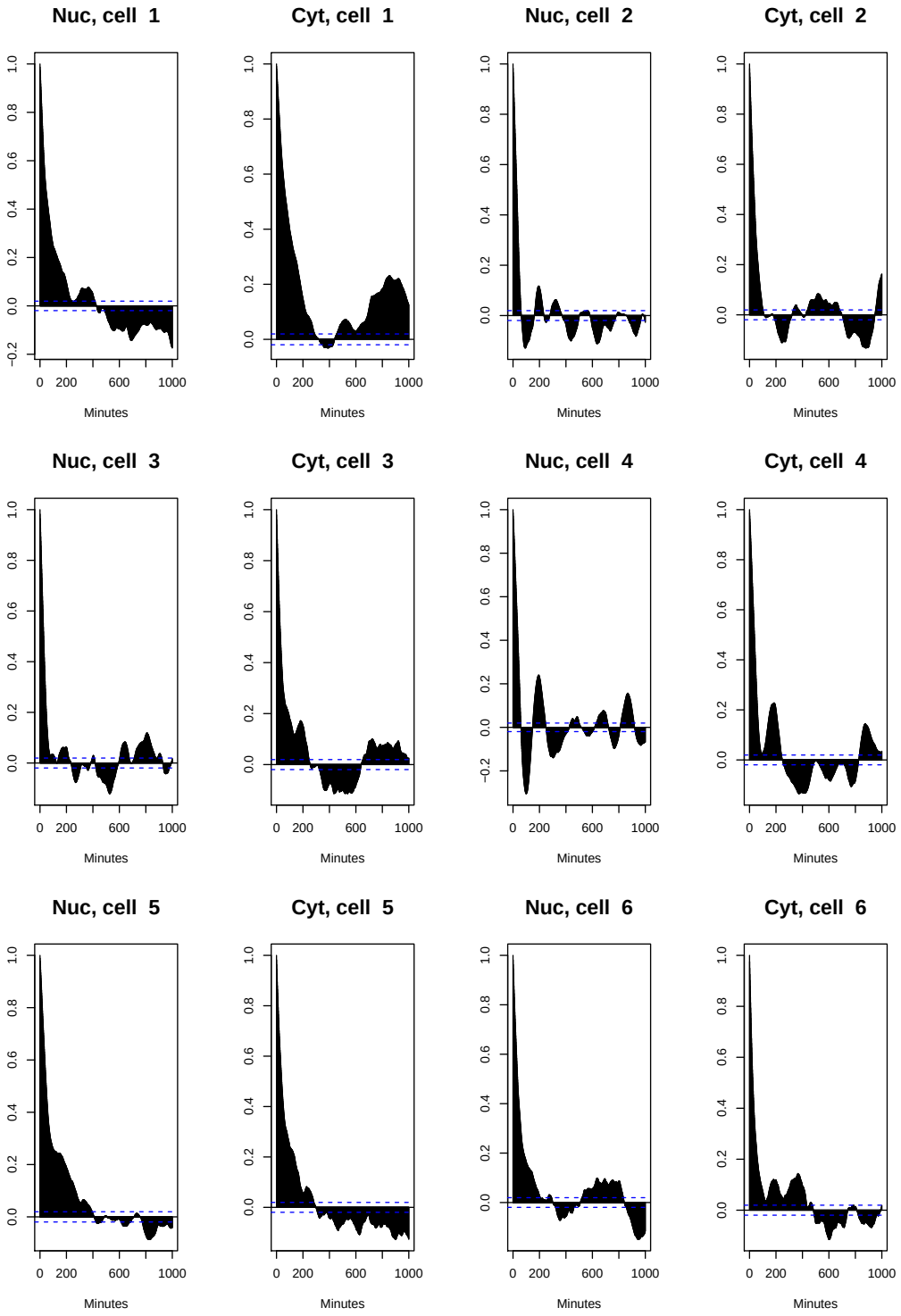


**Figure B.50:** Autocorrelation plots, for nucleus and cytoplasm, from Euler simulations perturbed with noise. Images refer to simulations from the posterior modes of parameters obtained from cells 1-6 under the basal condition.

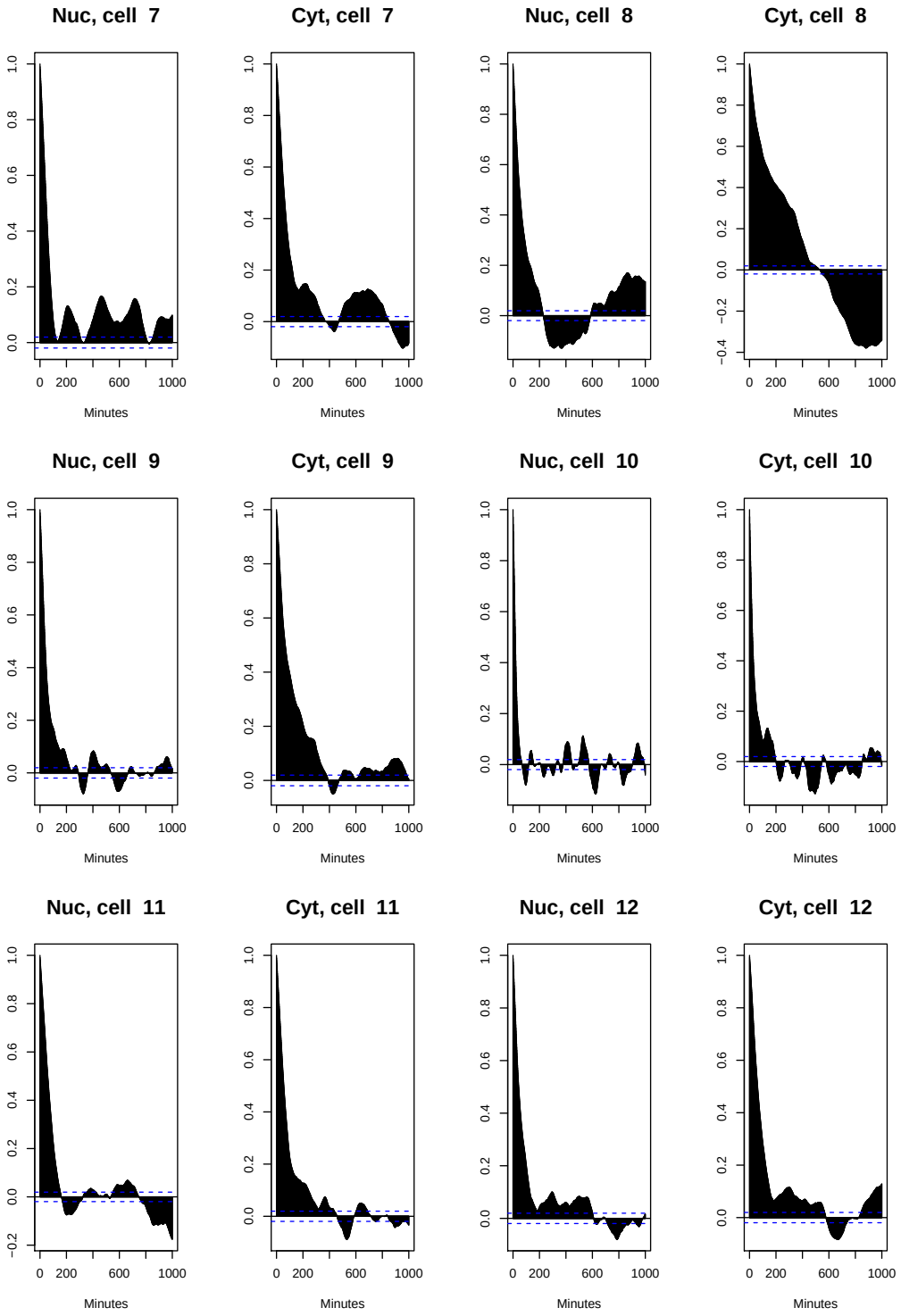


**Figure B.51:** Autocorrelation plots, for nucleus and cytoplasm, from Euler simulations perturbed with noise. Images refer to simulations from the posterior modes of parameters obtained from cells 7-12 under the basal condition.

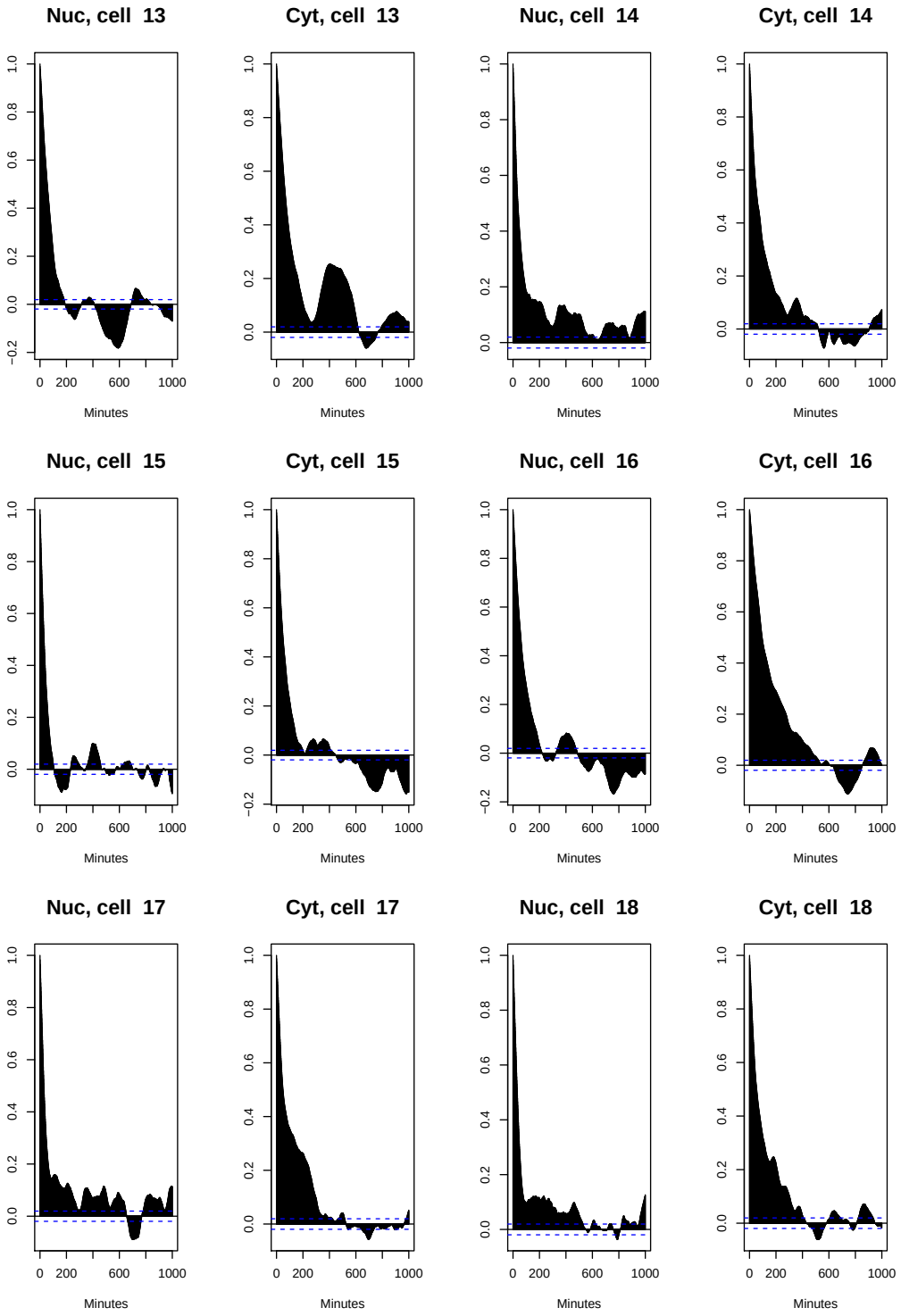


**Figure B.52:** Autocorrelation plots, for nucleus and cytoplasm, from Euler simulations perturbed with noise. Images refer to simulations from the posterior modes of parameters obtained from cells 13-18 under the basal condition.

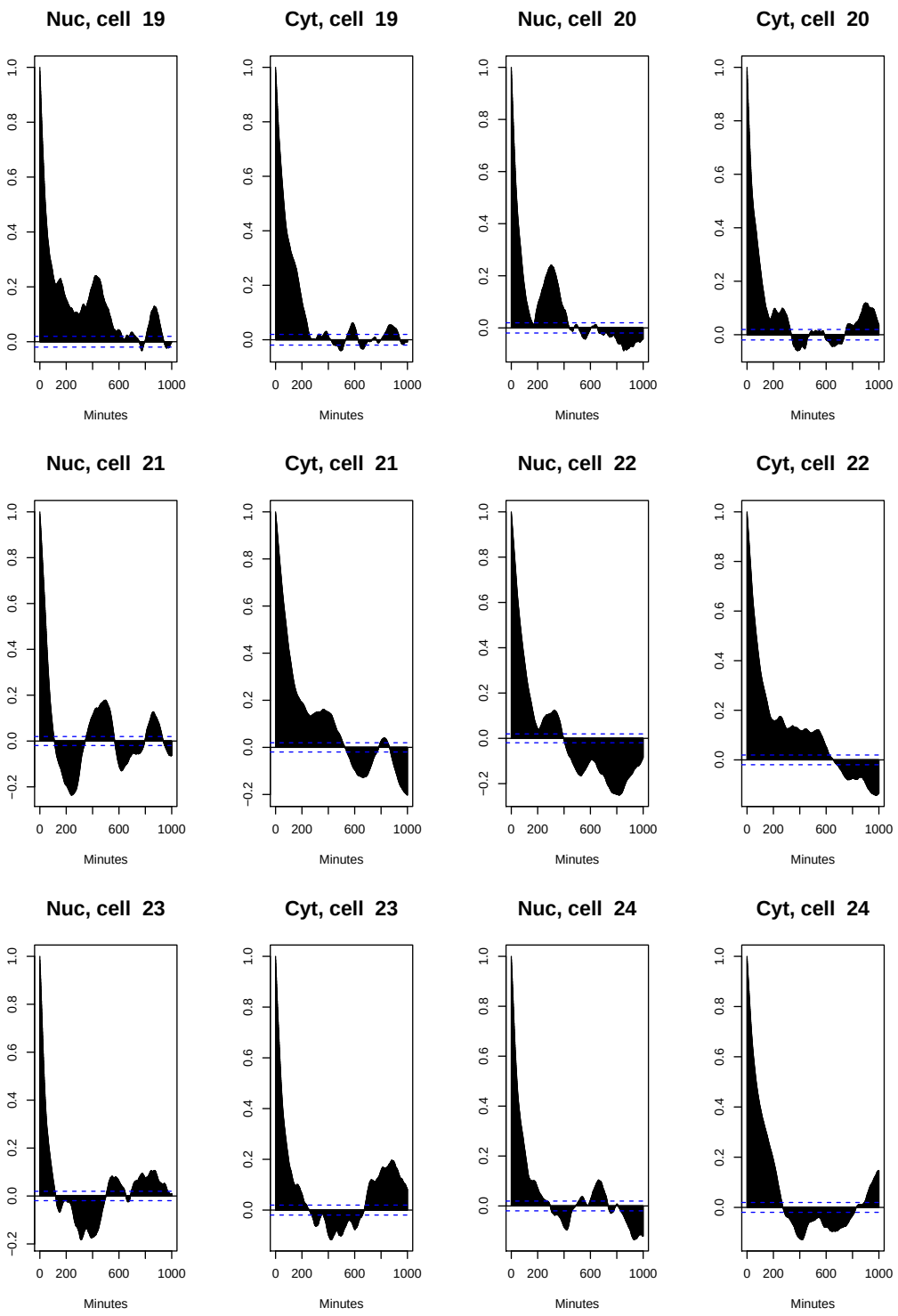


**Figure B.53:** Autocorrelation plots, for nucleus and cytoplasm, from Euler simulations perturbed with noise. Images refer to simulations from the posterior modes of parameters obtained from cells 19-24 under the basal condition.

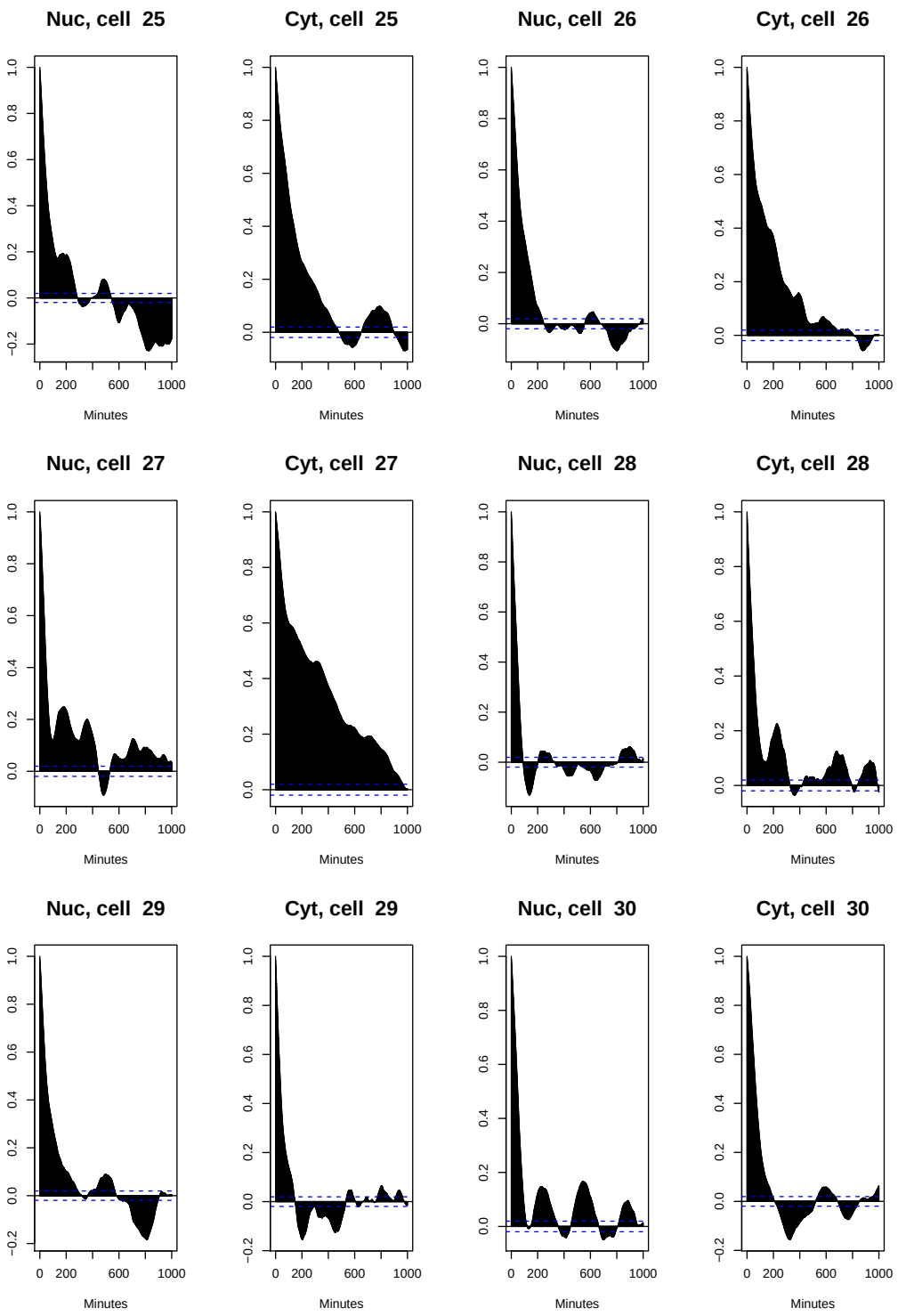


**Figure B.54:** Autocorrelation plots, for nucleus and cytoplasm, from Euler simulations perturbed with noise. Images refer to simulations from the posterior modes of parameters obtained from cells 25-30 under the basal condition.

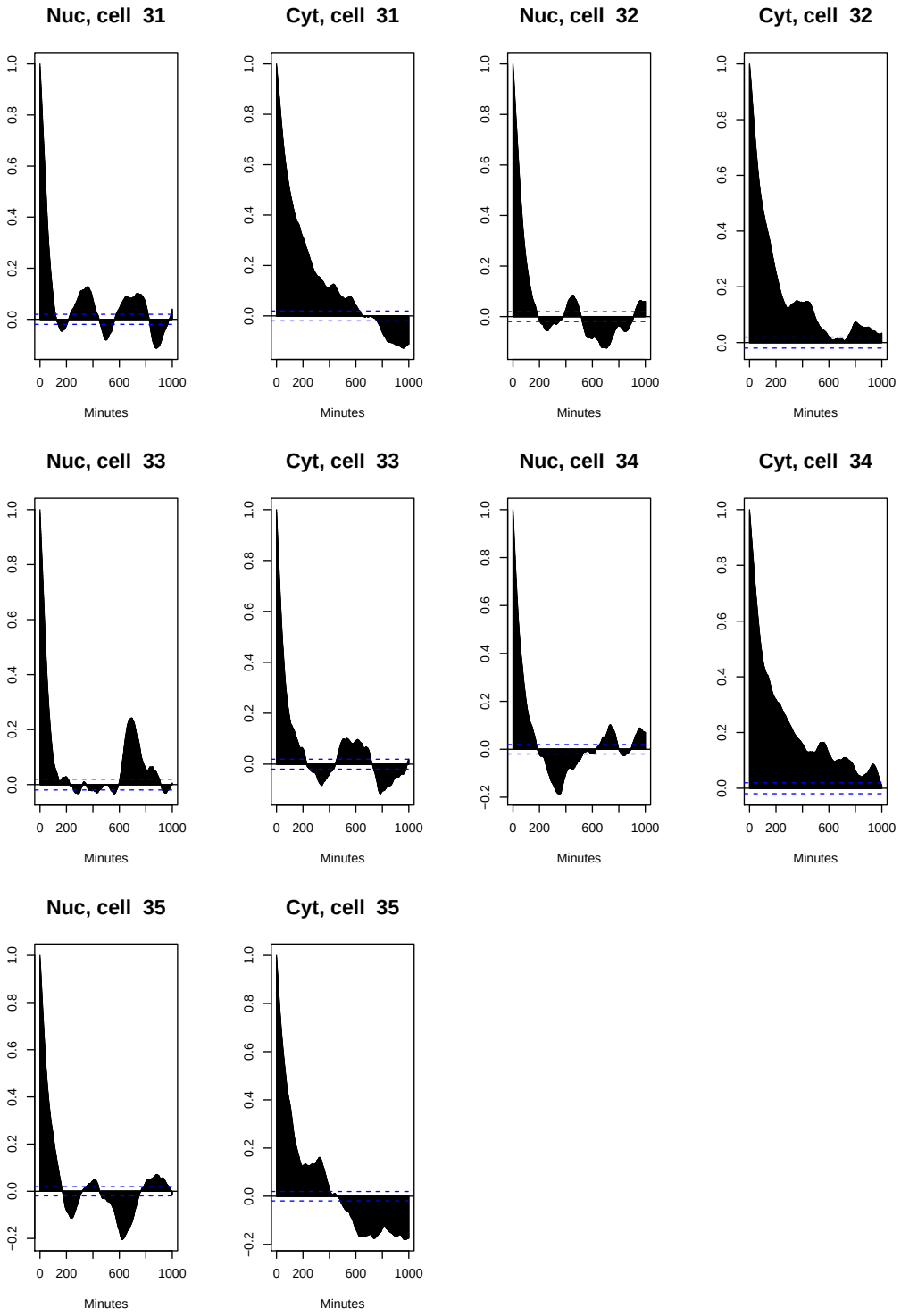


**Figure B.55:** Autocorrelation plots, for nucleus and cytoplasm, from Euler simulations perturbed with noise. Images refer to simulations from the posterior modes of parameters obtained from cells 31-35 under the basal condition.

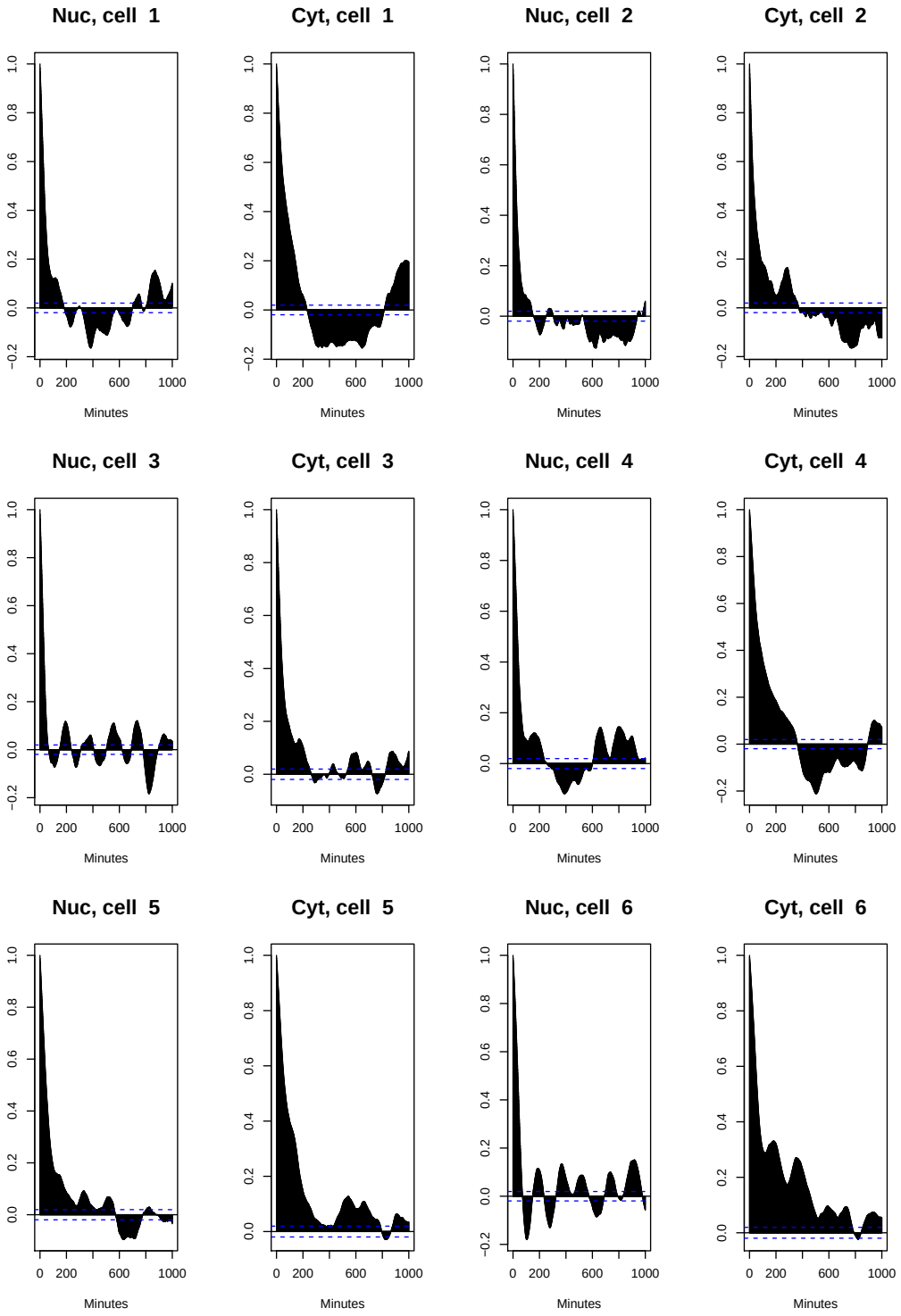


**Figure B.56:** Autocorrelation plots, for nucleus and cytoplasm, from Euler simulations perturbed with noise. Images refer to simulations from the posterior modes of parameters obtained from cells 1-6 under the stimulated condition.

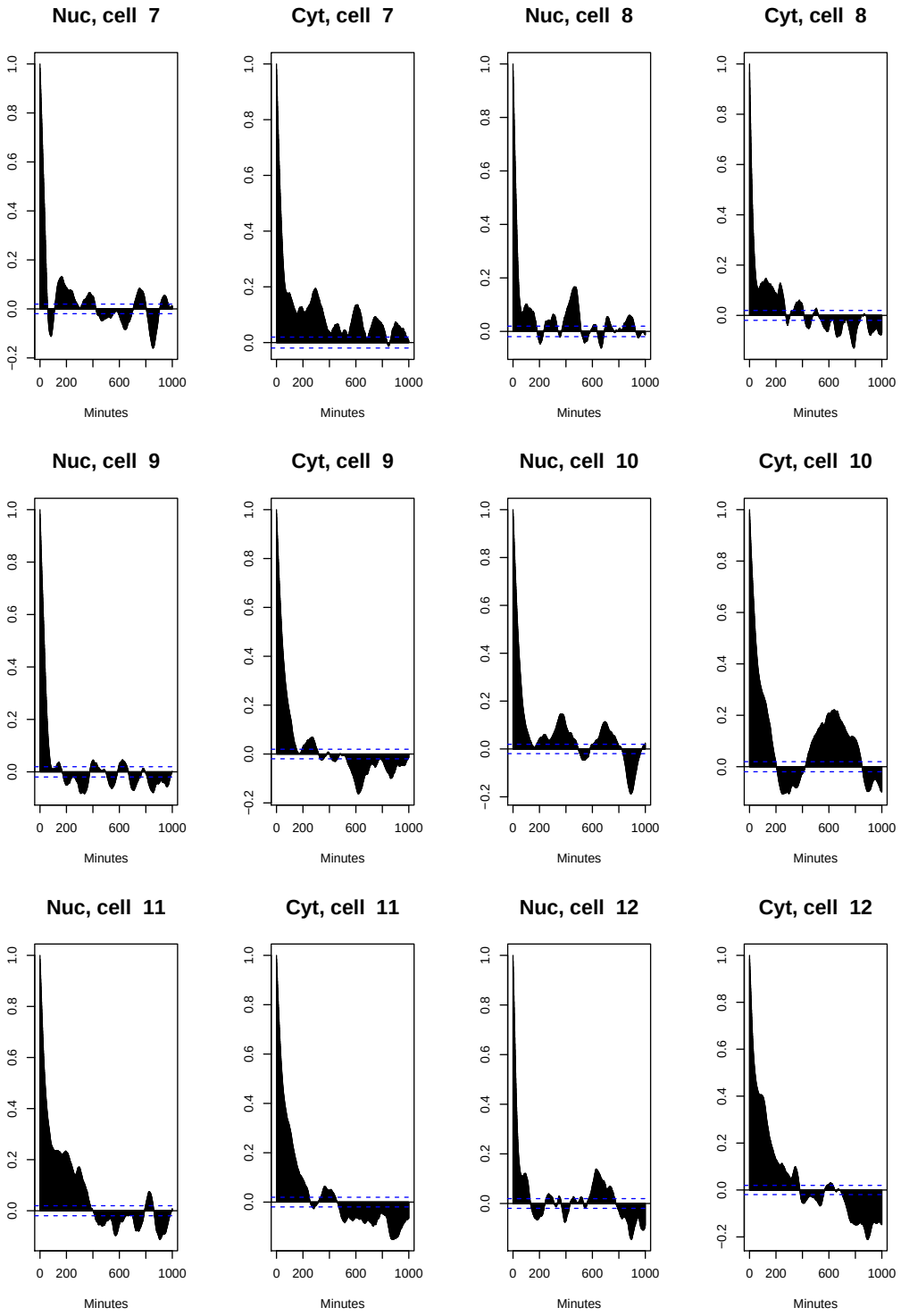


**Figure B.57:** Autocorrelation plots, for nucleus and cytoplasm, from Euler simulations perturbed with noise. Images refer to simulations from the posterior modes of parameters obtained from cells 7-12 under the stimulated condition.

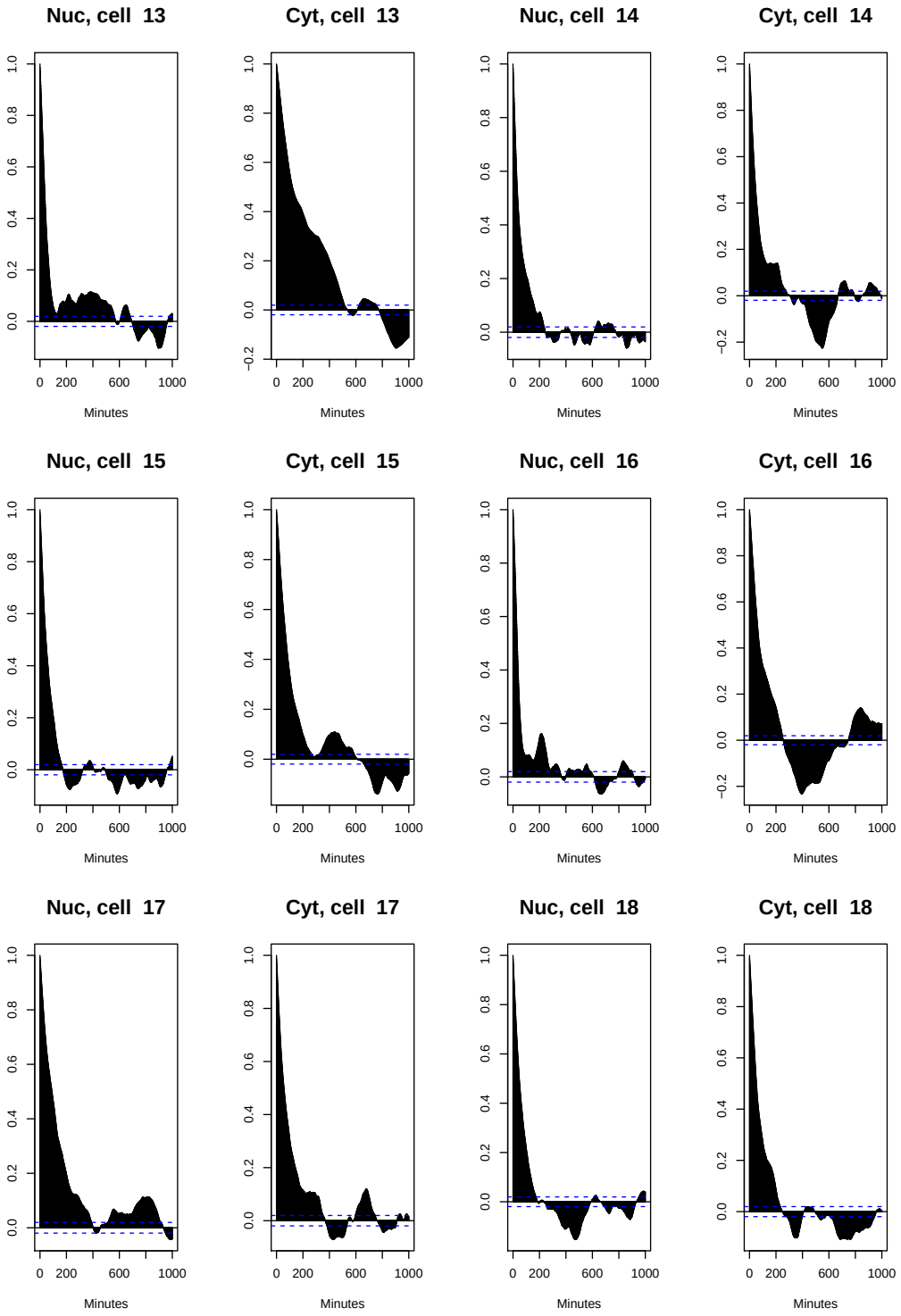


**Figure B.58:** Autocorrelation plots, for nucleus and cytoplasm, from Euler simulations perturbed with noise. Images refer to simulations from the posterior modes of parameters obtained from cells 13-18 under the stimulated condition.

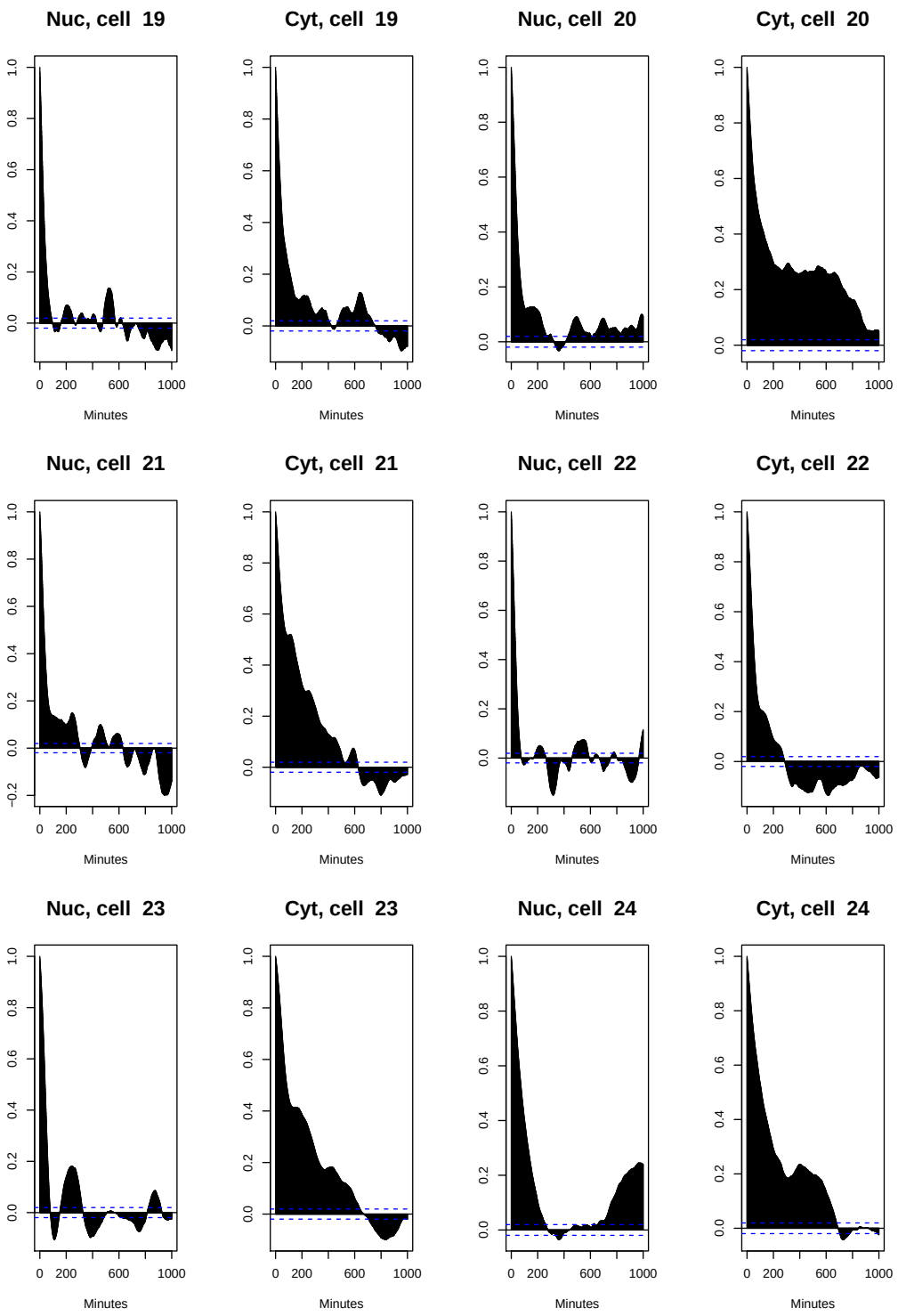


**Figure B.59:** Autocorrelation plots, for nucleus and cytoplasm, from Euler simulations perturbed with noise. Images refer to simulations from the posterior modes of parameters obtained from cells 19-24 under the stimulated condition.

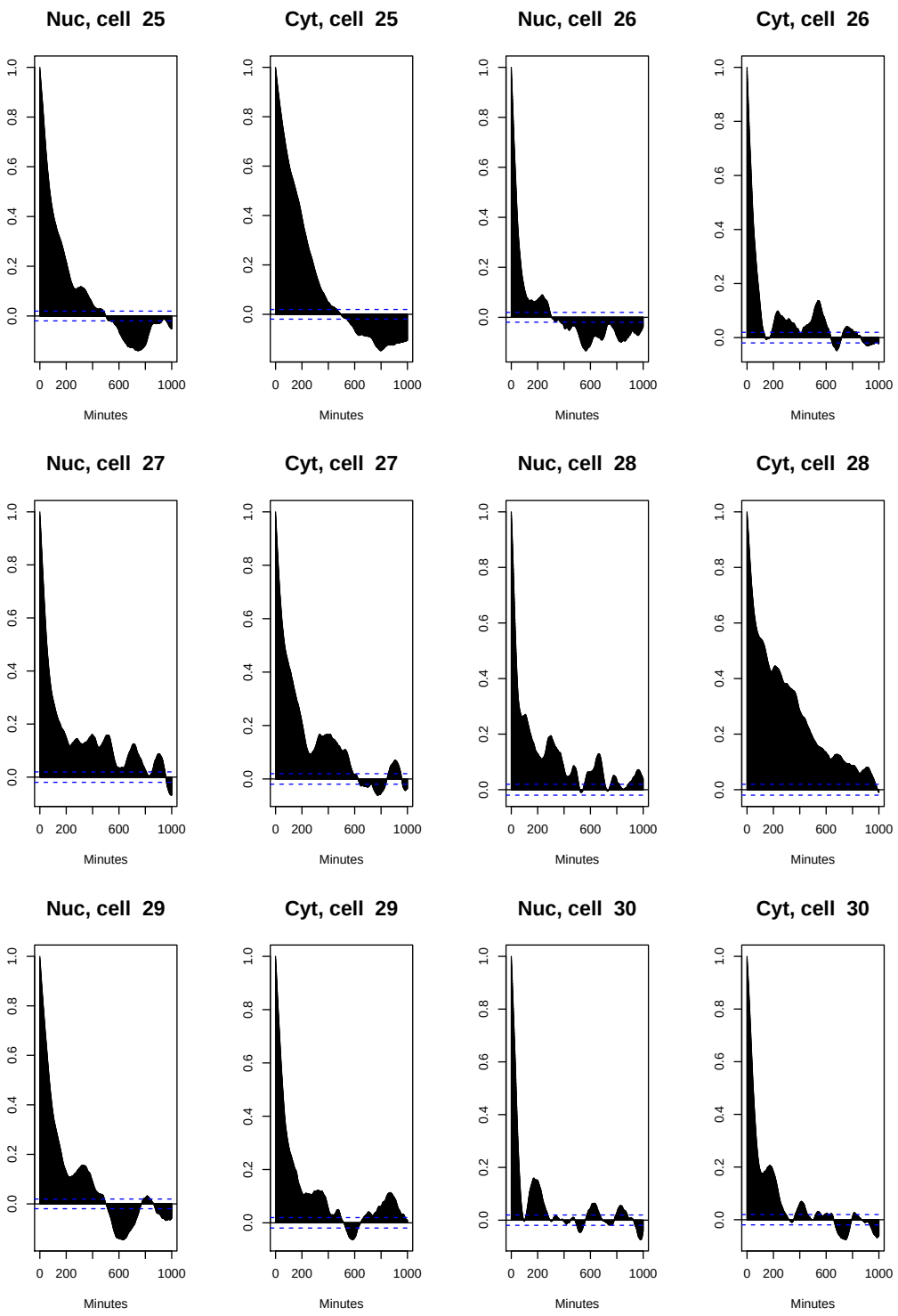


**Figure B.60:** Autocorrelation plots, for nucleus and cytoplasm, from Euler simulations perturbed with noise. Images refer to simulations from the posterior modes of parameters obtained from cells 25-30 under the stimulated condition.

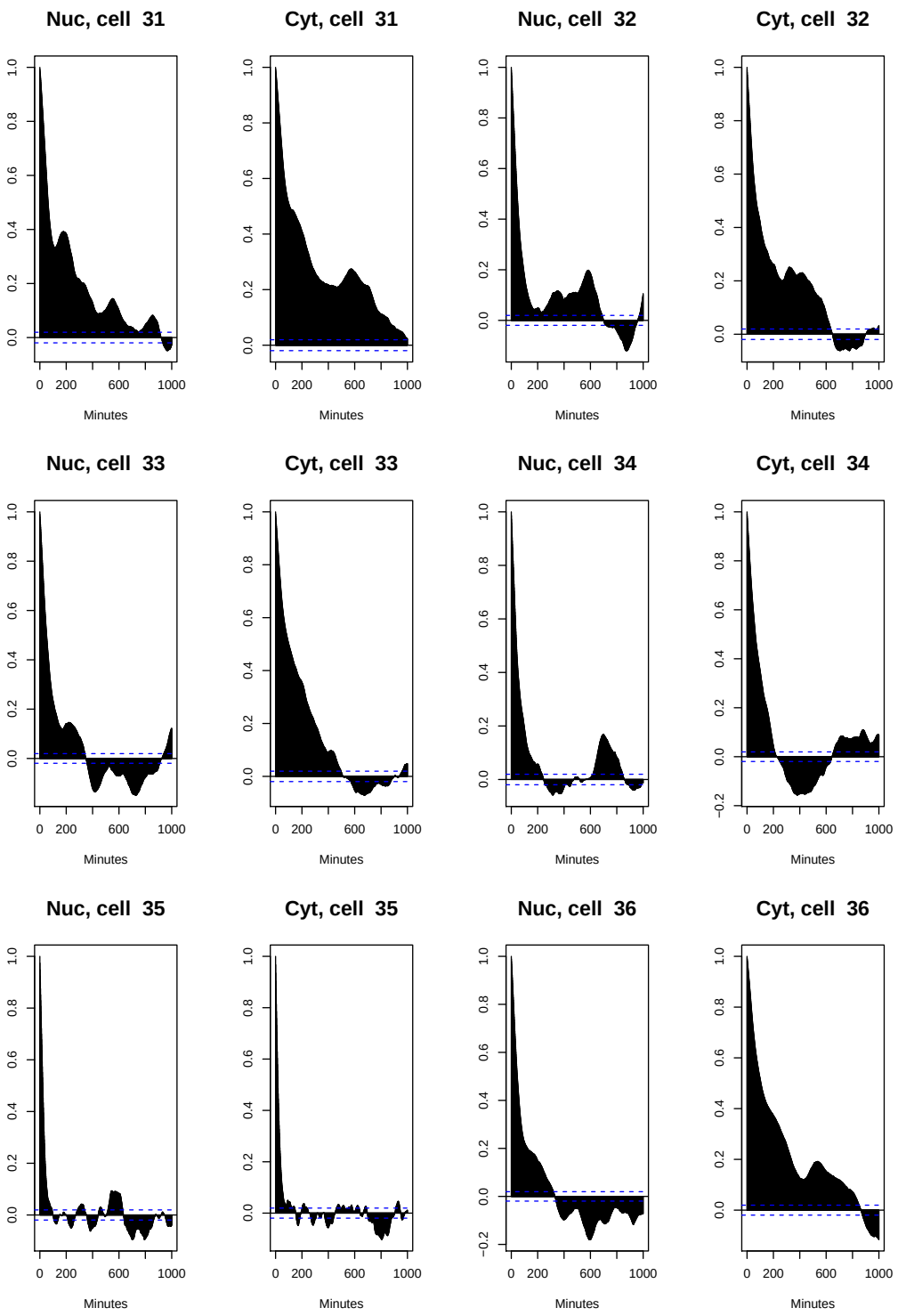


**Figure B.61:** Autocorrelation plots, for nucleus and cytoplasm, from Euler simulations perturbed with noise. Images refer to simulations from the posterior modes of parameters obtained from cells 31-36 under the stimulated condition.

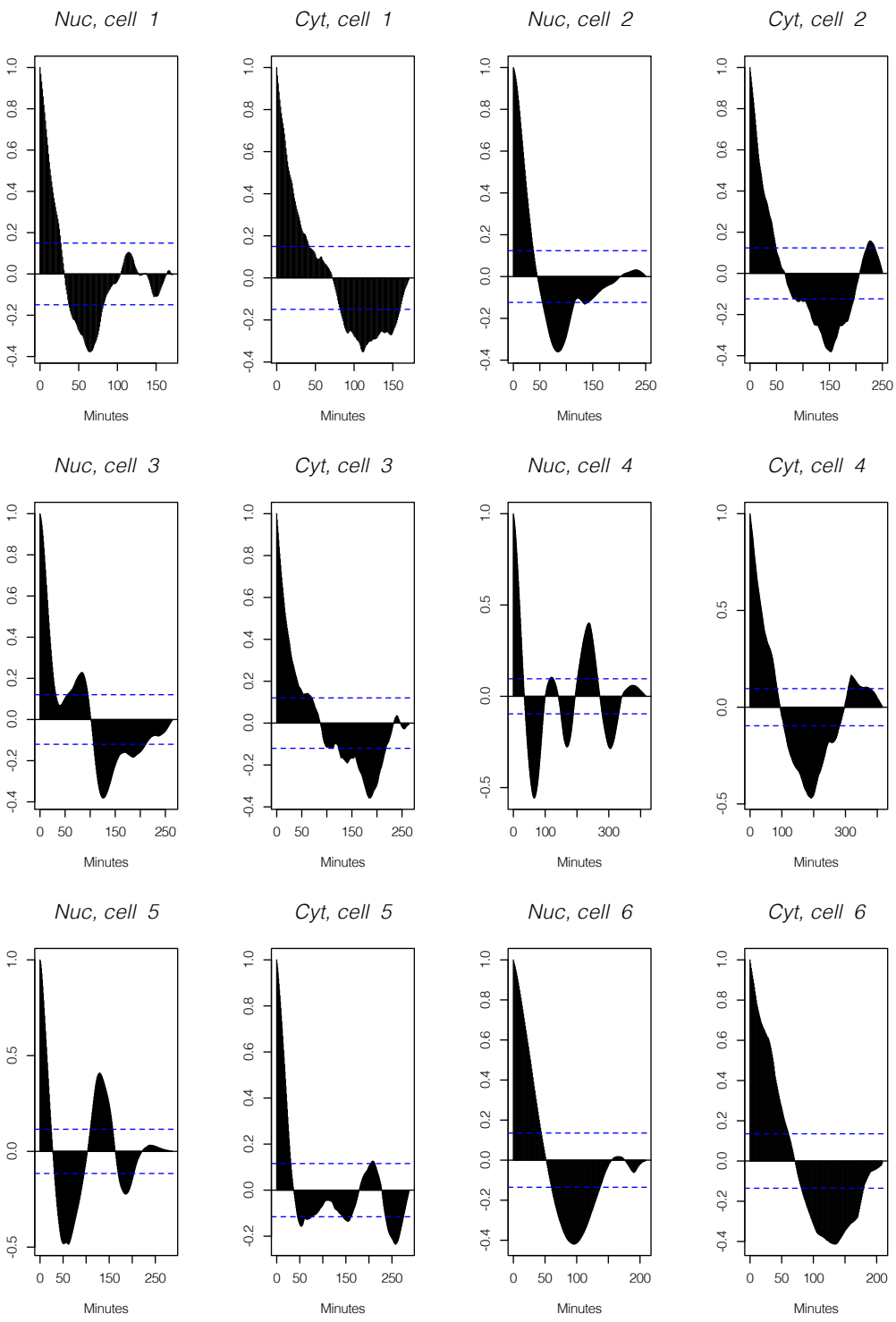


**Figure B.62:** Empirical autocorrelation plots, for nucleus and cytoplasm, from experimental data for cells 1-6 under the basal condition.

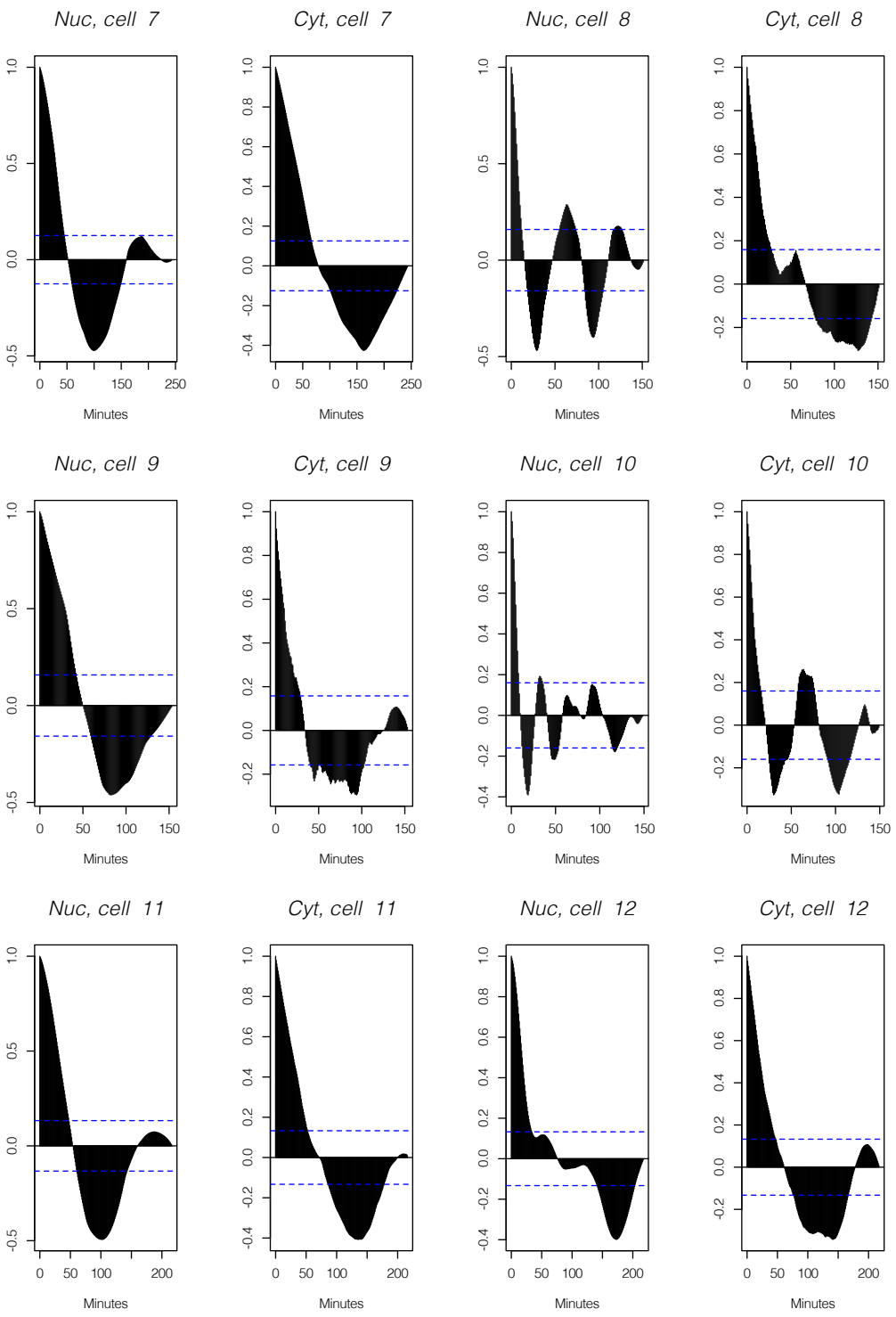


**Figure B.63:** Empirical autocorrelation plots, for nucleus and cytoplasm, from experimental data for cells 7-12 under the basal condition.

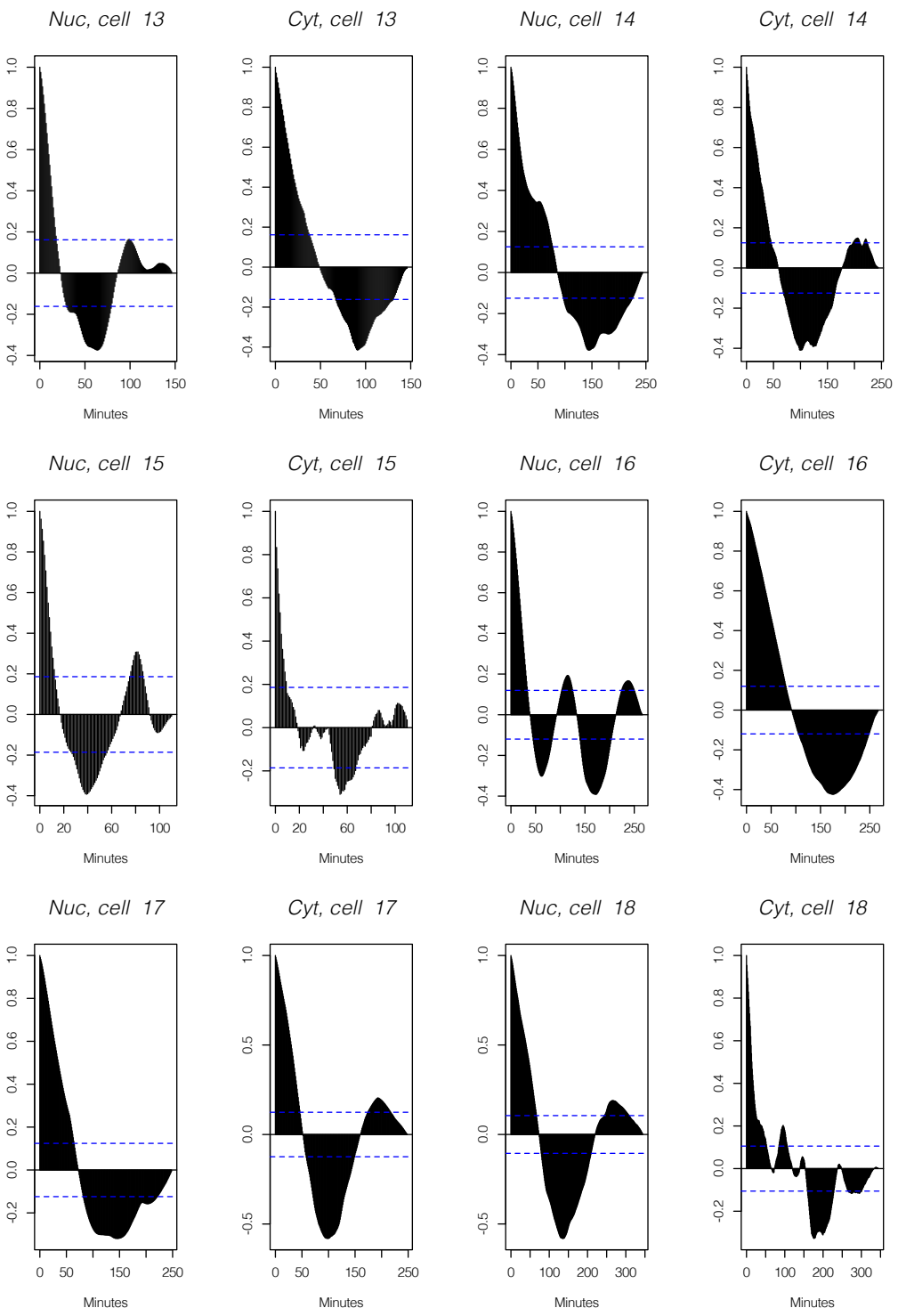


**Figure B.64:** Empirical autocorrelation plots, for nucleus and cytoplasm, from experimental data for cells 13-18 under the basal condition.

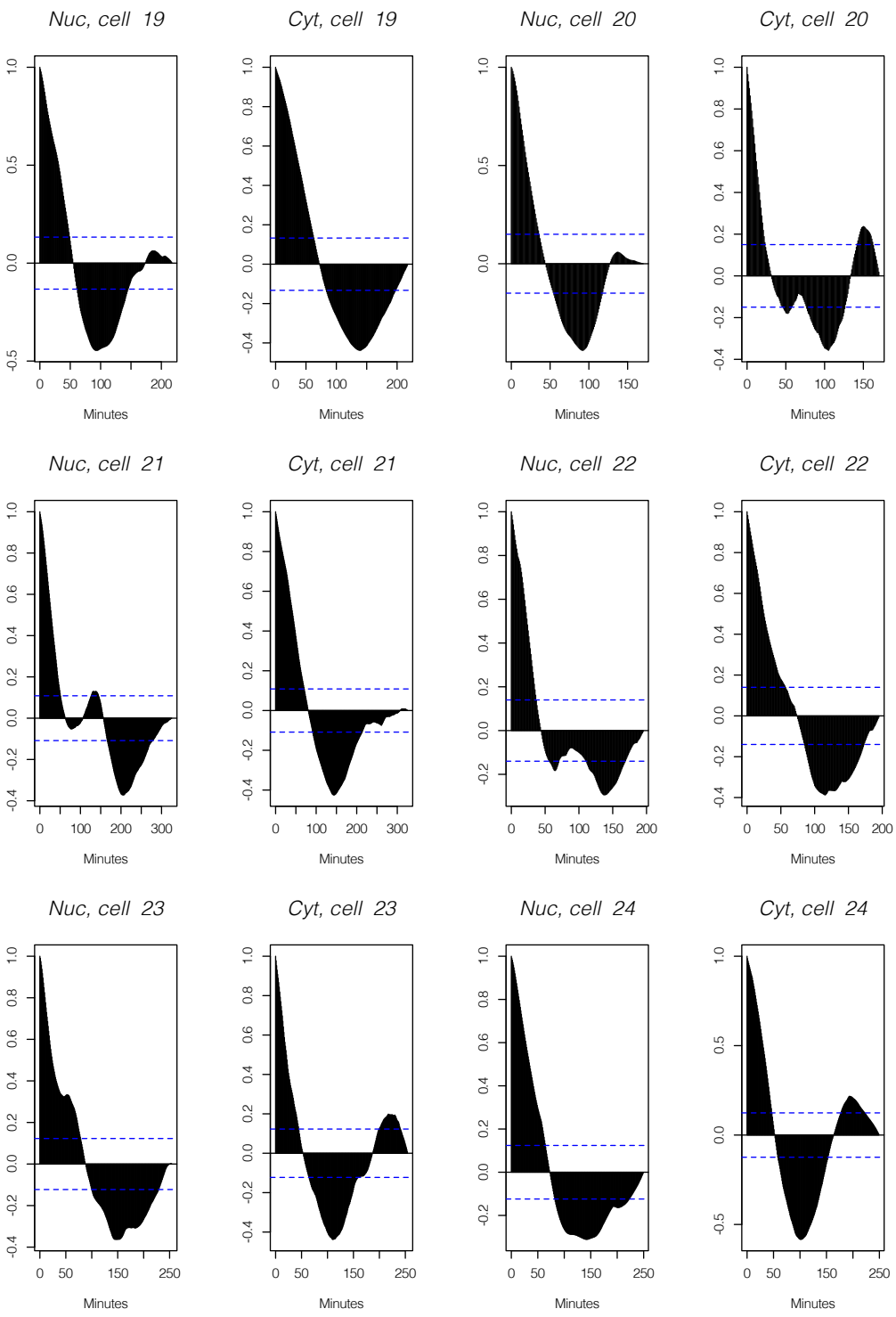


**Figure B.65:** Empirical autocorrelation plots, for nucleus and cytoplasm, from experimental data for cells 19-24 under the basal condition.

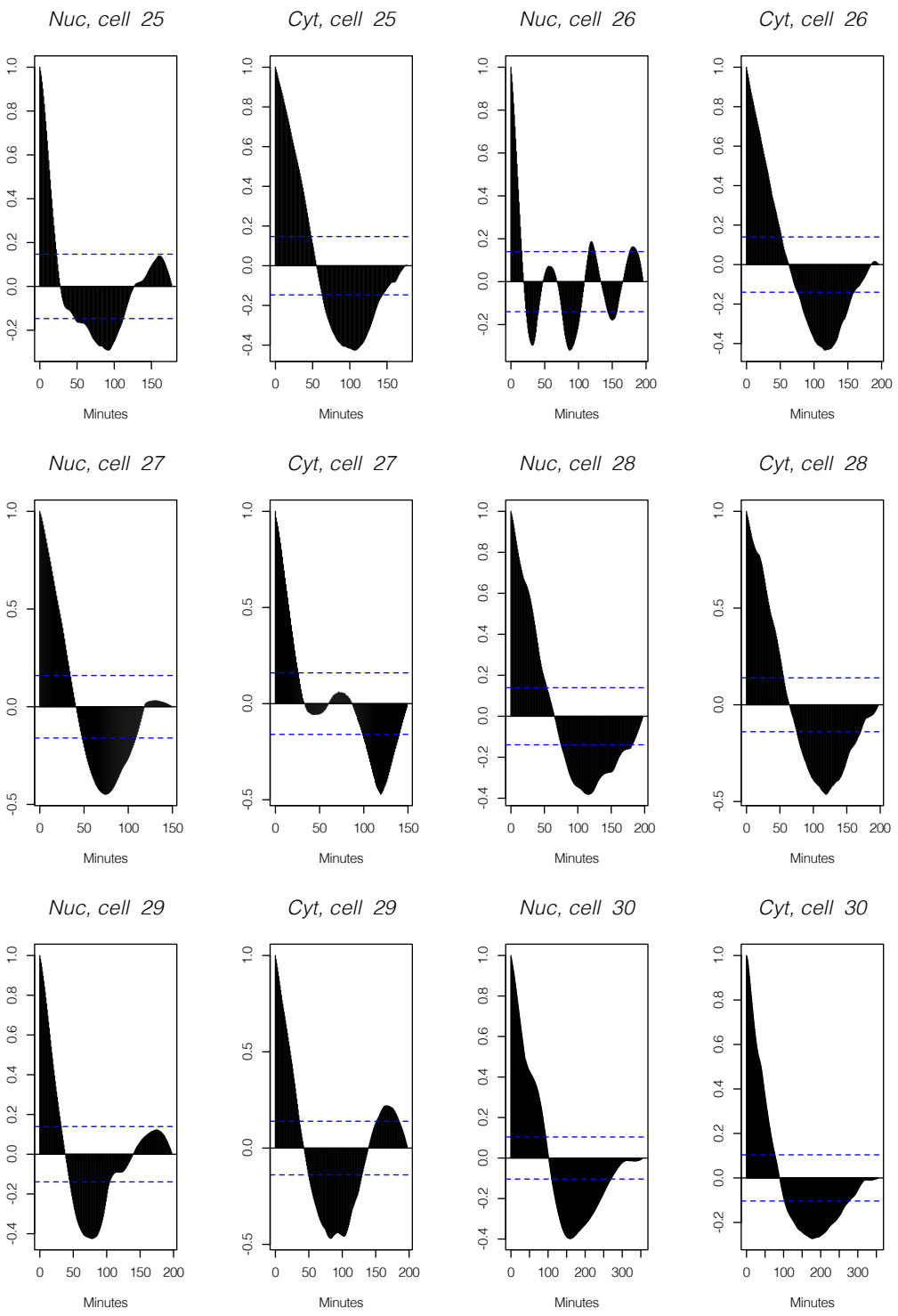


**Figure B.66:** Empirical autocorrelation plots, for nucleus and cytoplasm, from experimental data for cells 25-30 under the basal condition.

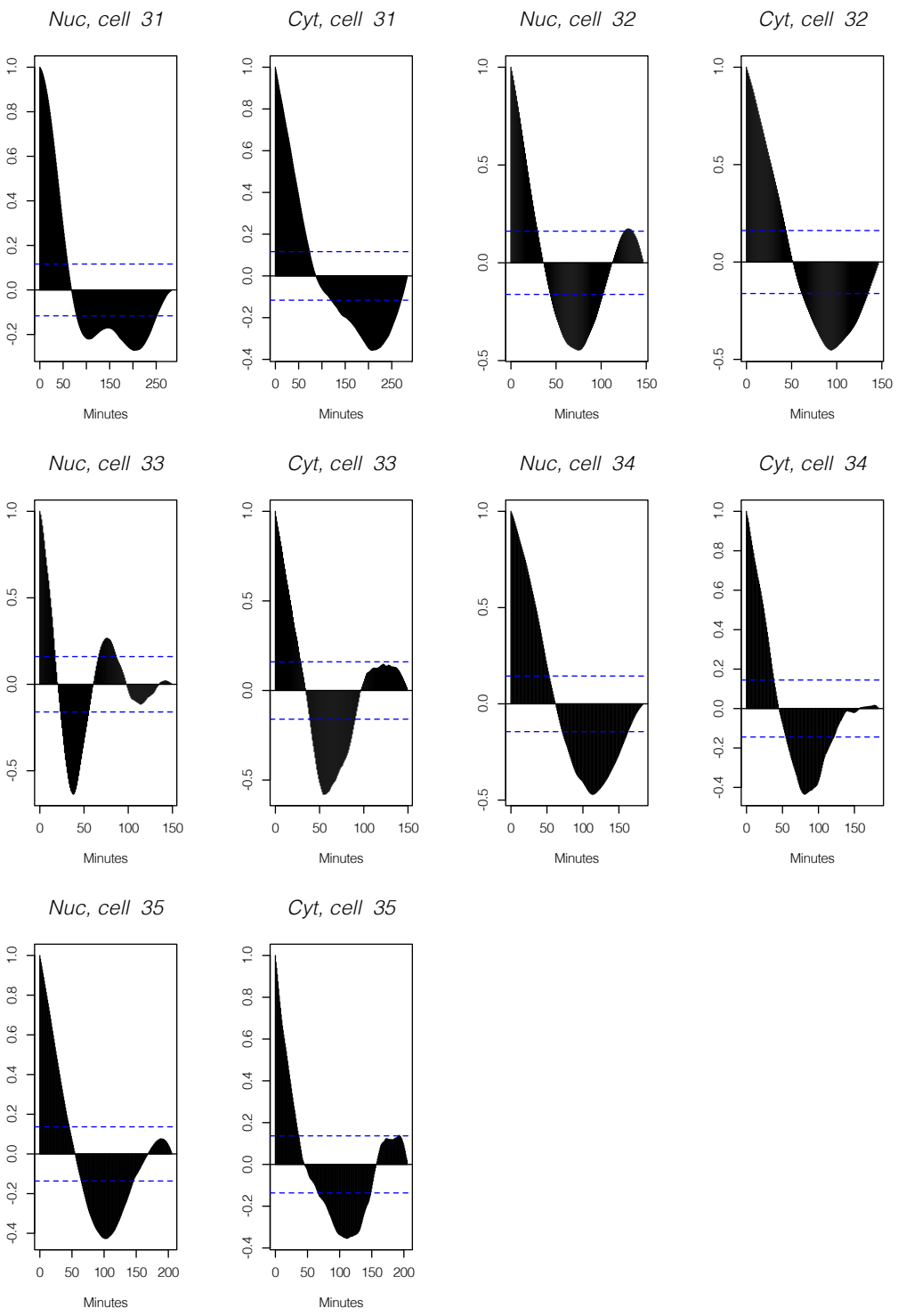


**Figure B.67:** Empirical autocorrelation plots, for nucleus and cytoplasm, from experimental data for cells 31-35 under the basal condition.

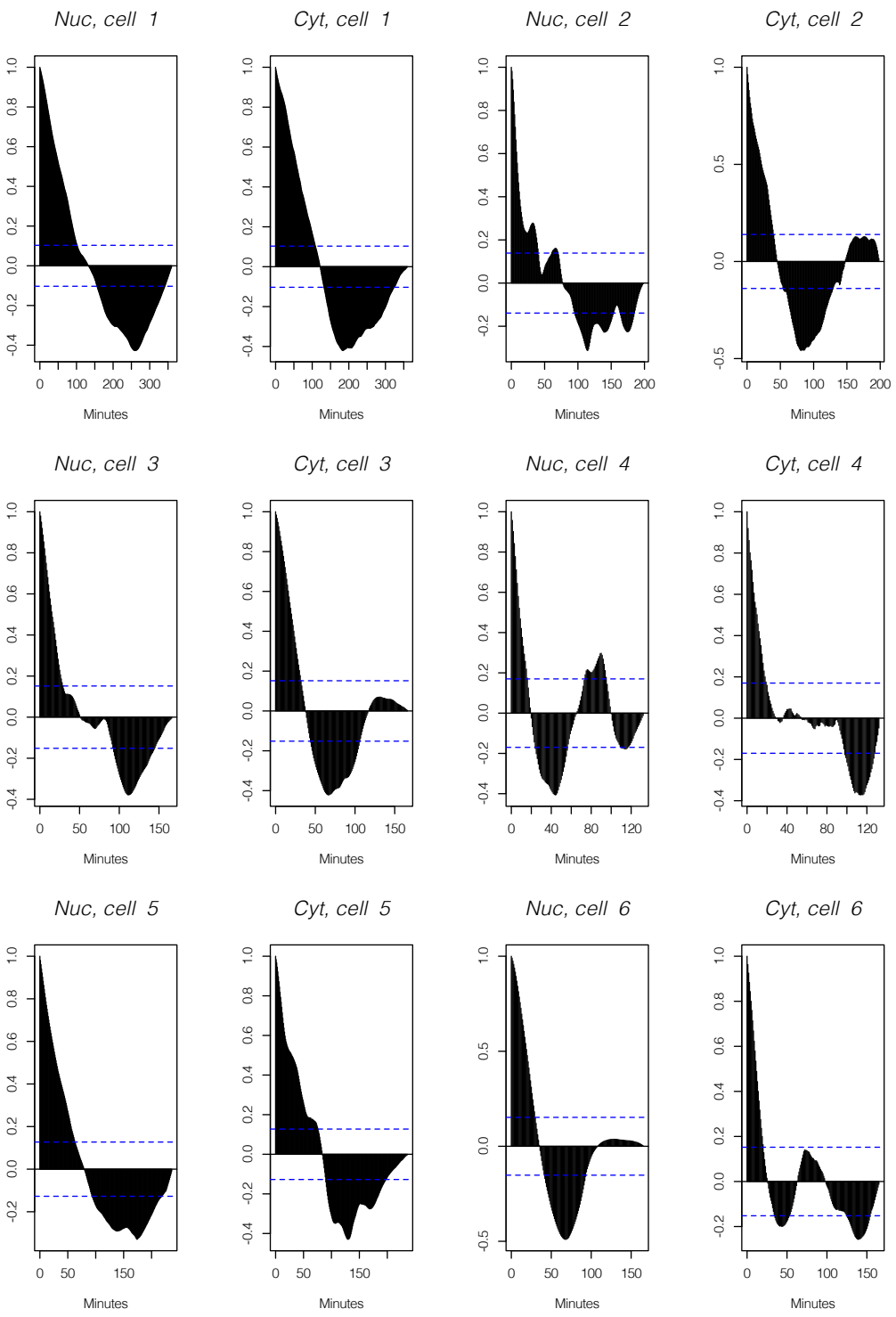


**Figure B.68:** Empirical autocorrelation plots, for nucleus and cytoplasm, from experimental data for cells 1-6 under the stimulated condition.

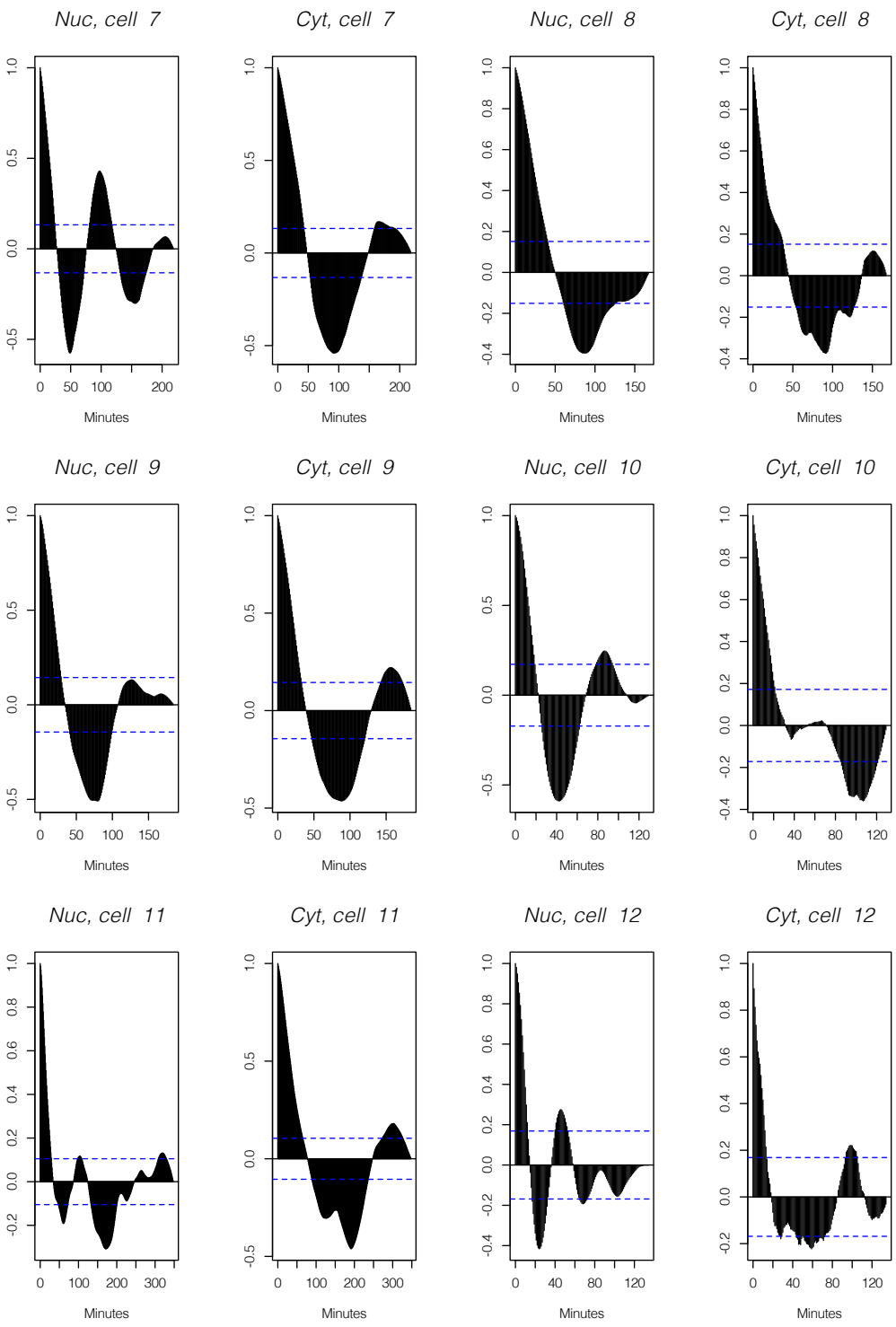


**Figure B.69:** Empirical autocorrelation plots, for nucleus and cytoplasm, from experimental data for cells 7-12 under the stimulated condition.

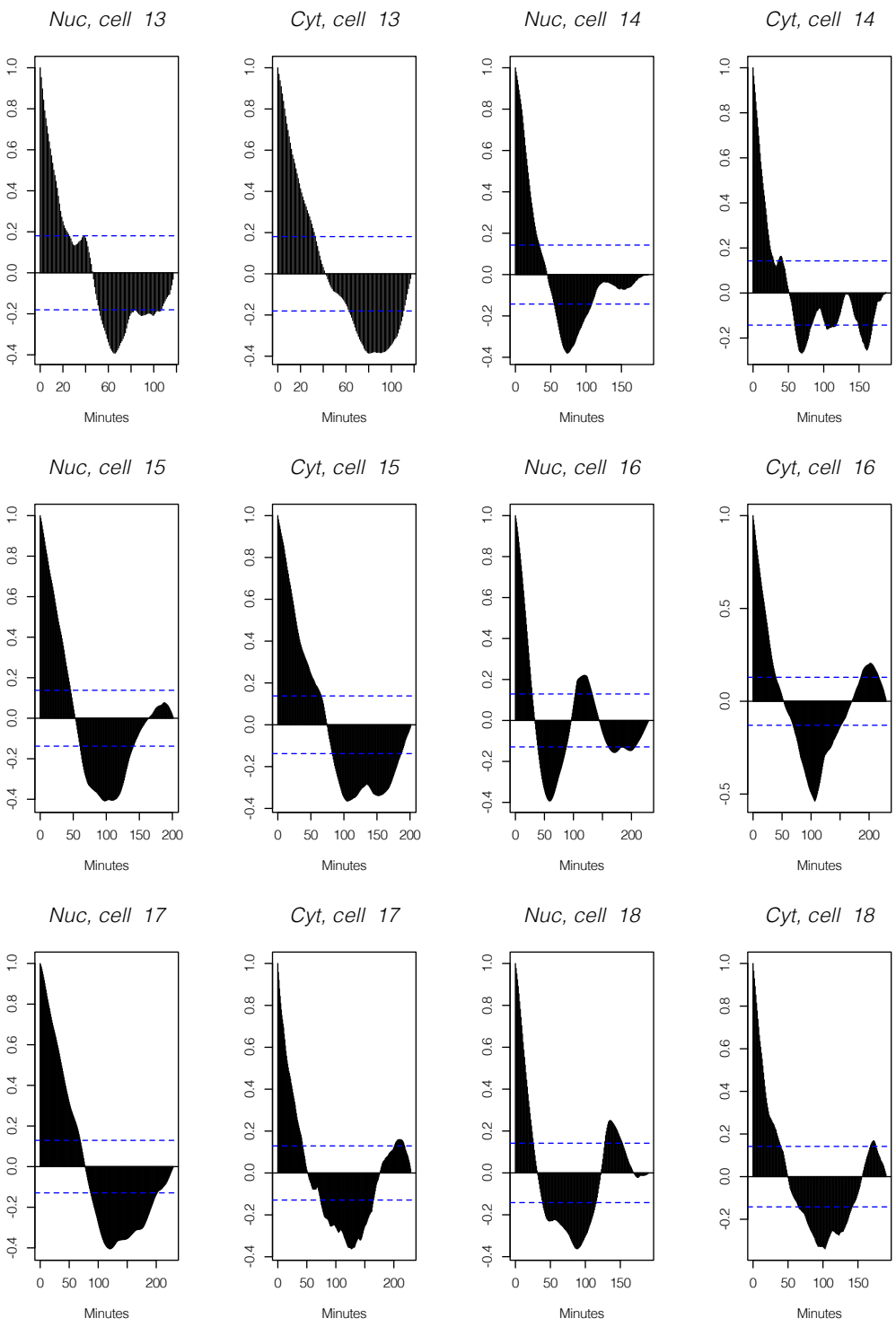


**Figure B.70:** Empirical autocorrelation plots, for nucleus and cytoplasm, from experimental data for cells 13-18 under the stimulated condition.

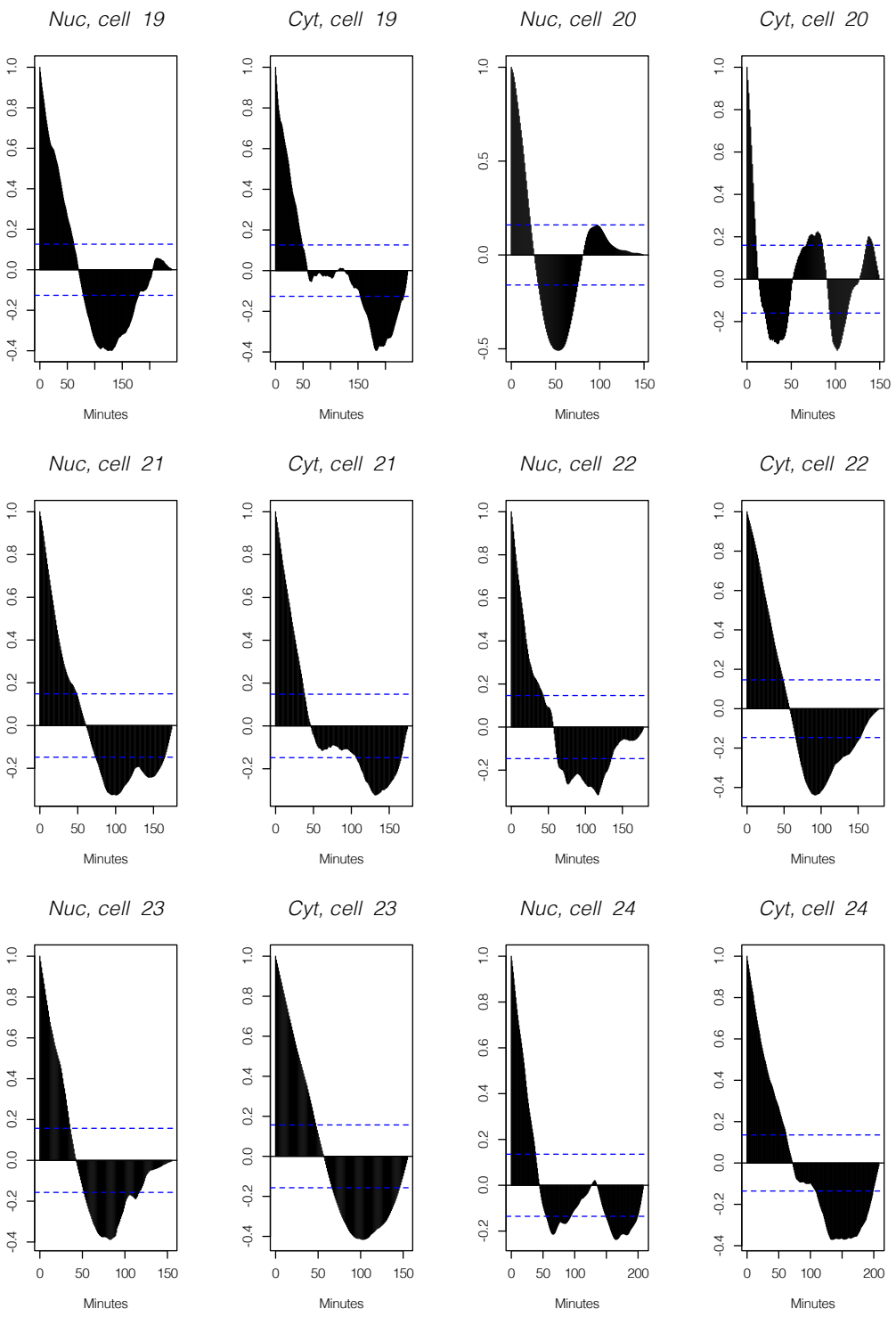


**Figure B.71:** Empirical autocorrelation plots, for nucleus and cytoplasm, from experimental data for cells 19-24 under the stimulated condition.

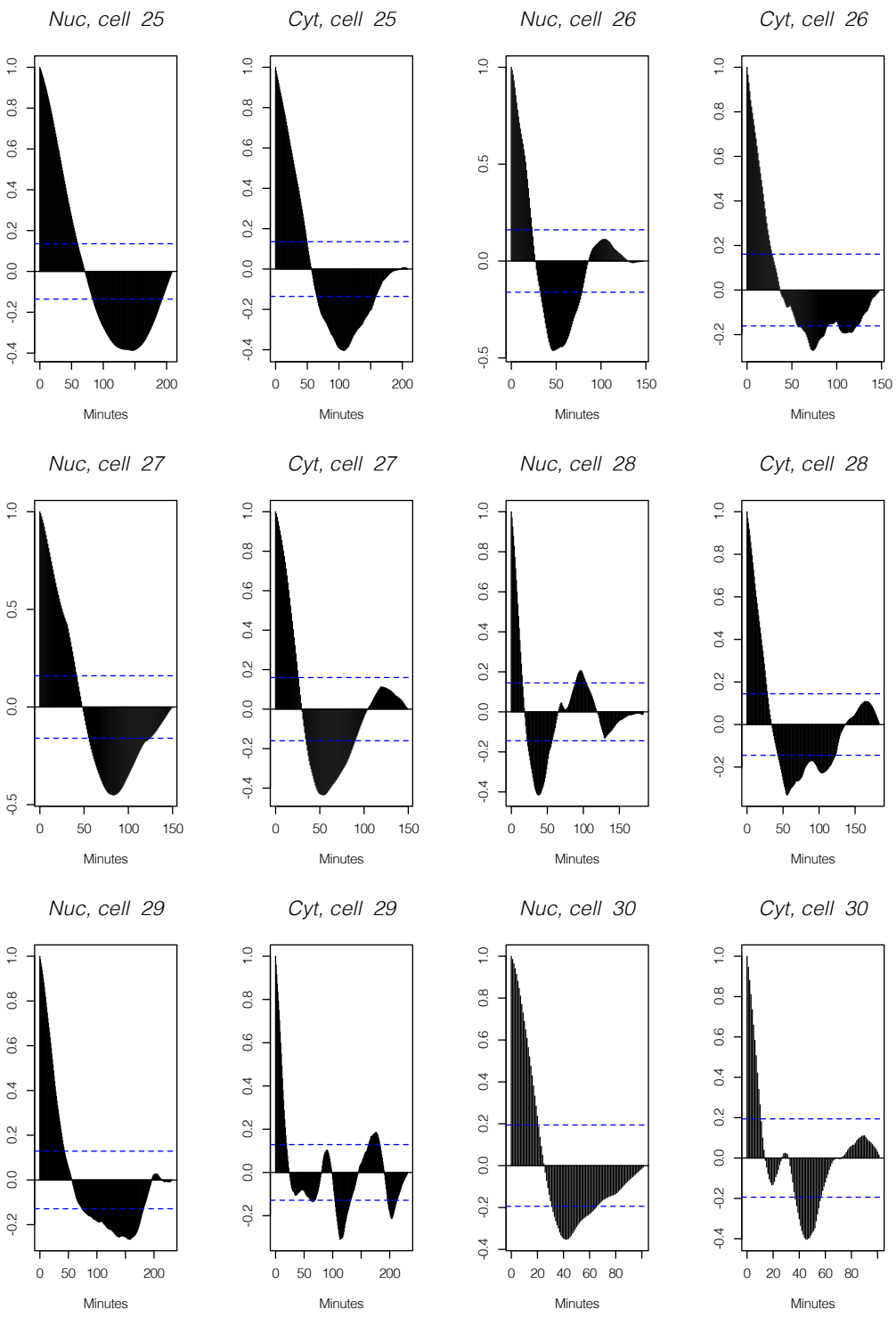


**Figure B.72:** Empirical autocorrelation plots, for nucleus and cytoplasm, from experimental data for cells 25-30 under the stimulated condition.

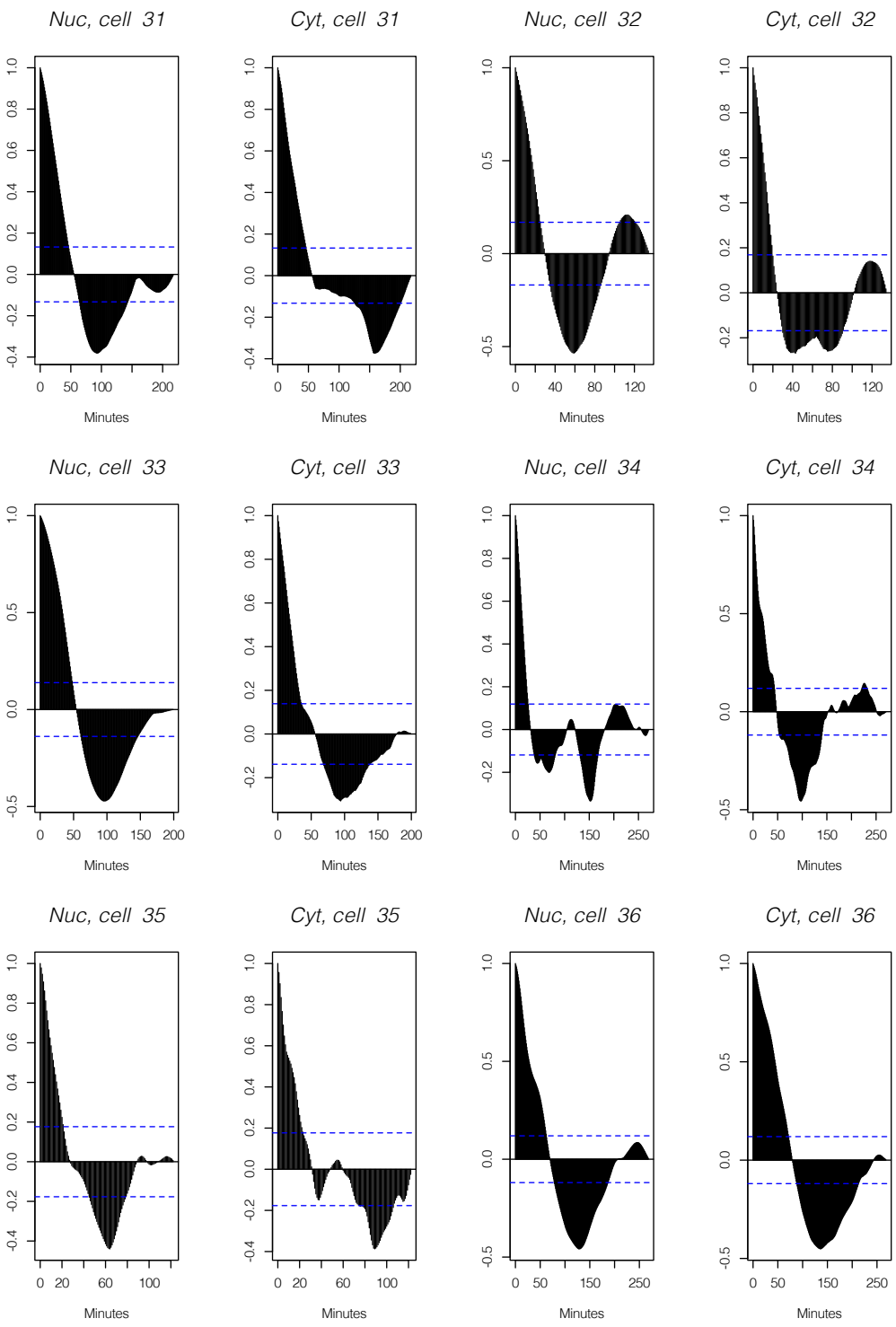


**Figure B.73:** Empirical autocorrelation plots, for nucleus and cytoplasm, from experimental data for cells 31-36 under the stimulated condition.

## B.2 Part II

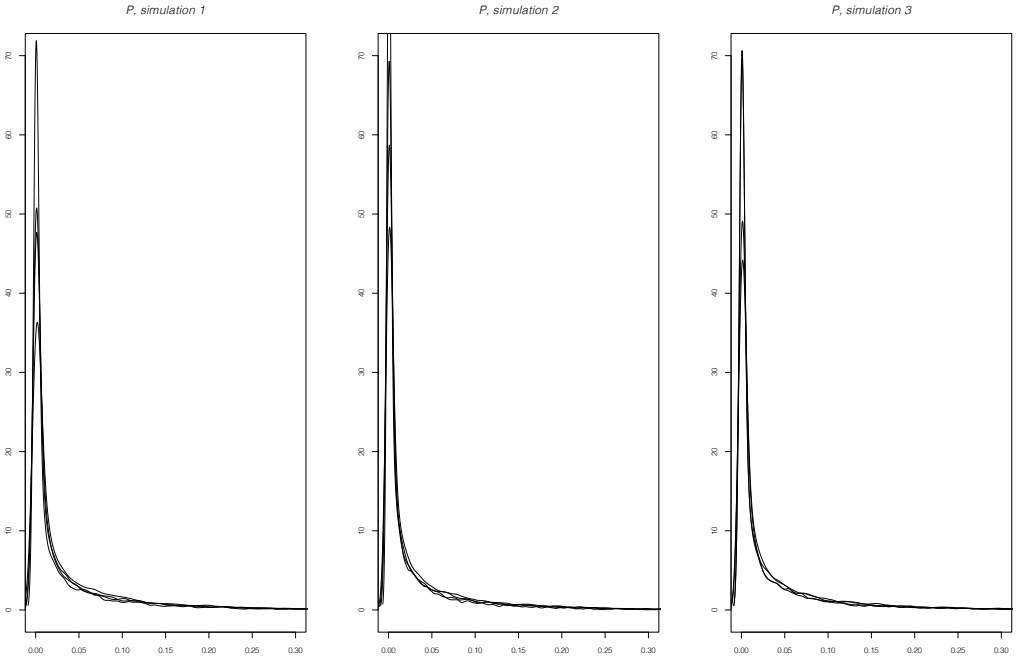


**Figure B.74:** Densities of the simulated latent states in $P^{(k)}$ from three simulation studies. Each image refers to the four replicates of a simulation.

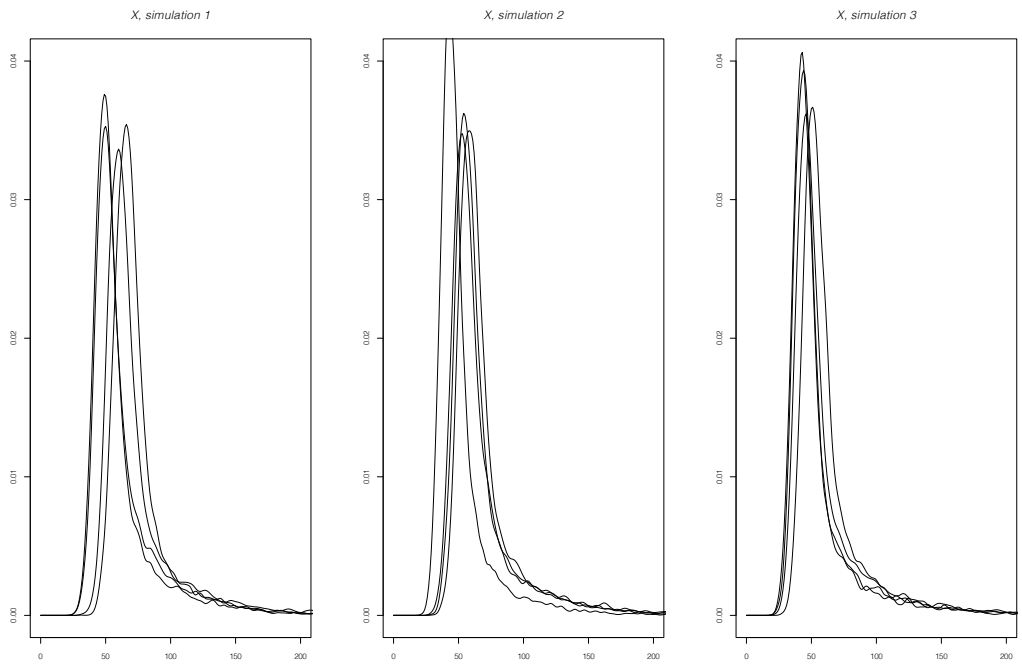


**Figure B.75:** Densities of the simulated latent states in $X^{(k)}$ from three simulation studies. Each image refers to the four replicates of a simulation.

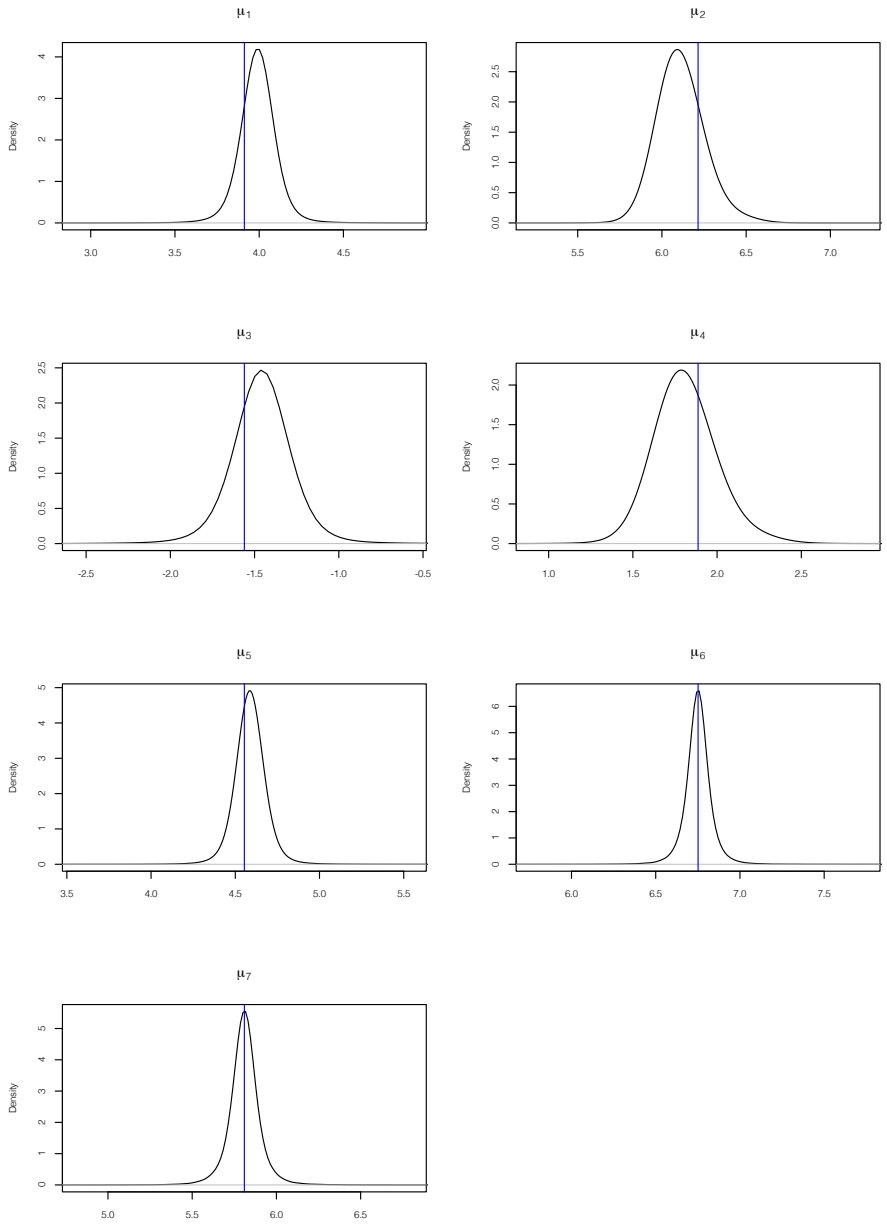


**Figure B.76:** Posterior densities for the hypermean parameters from the first simulation study. The vertical line denotes the real value used in the simulation. Images, from left to right and top to bottom, refer in the order to the hypermeans of $\tilde{\alpha}_0^{(k)}, \tilde{\alpha}_1^{(k)}, \tilde{k}_{ON}^{(k)}, \tilde{k}_{OFF}^{(k)}, \kappa^{(k)}, \mu_\epsilon^{(k)}$ and $\sigma_\epsilon^{(k)}$. Note that, since we did not simulate the hierarchical parameters $\mu_\epsilon^{(k)}$ and $\sigma_\epsilon^{(k)}$ from a hyperprior distribution, the lines for these two parameters refer to the mean of the respective logarithmic hierarchical parameters over the four replicates, i.e. $\frac{1}{4}\left(log(\mu_\epsilon^{(1)}) + log(\mu_\epsilon^{(2)}) + log(\mu_\epsilon^{(3)}) + log(\mu_\epsilon^{(4)})\right)$ and $\frac{1}{4}\left(log(\sigma_\epsilon^{(1)}) + log(\sigma_\epsilon^{(2)}) + log(\sigma_\epsilon^{(3)}) + log(\sigma_\epsilon^{(4)})\right)$, respectively.

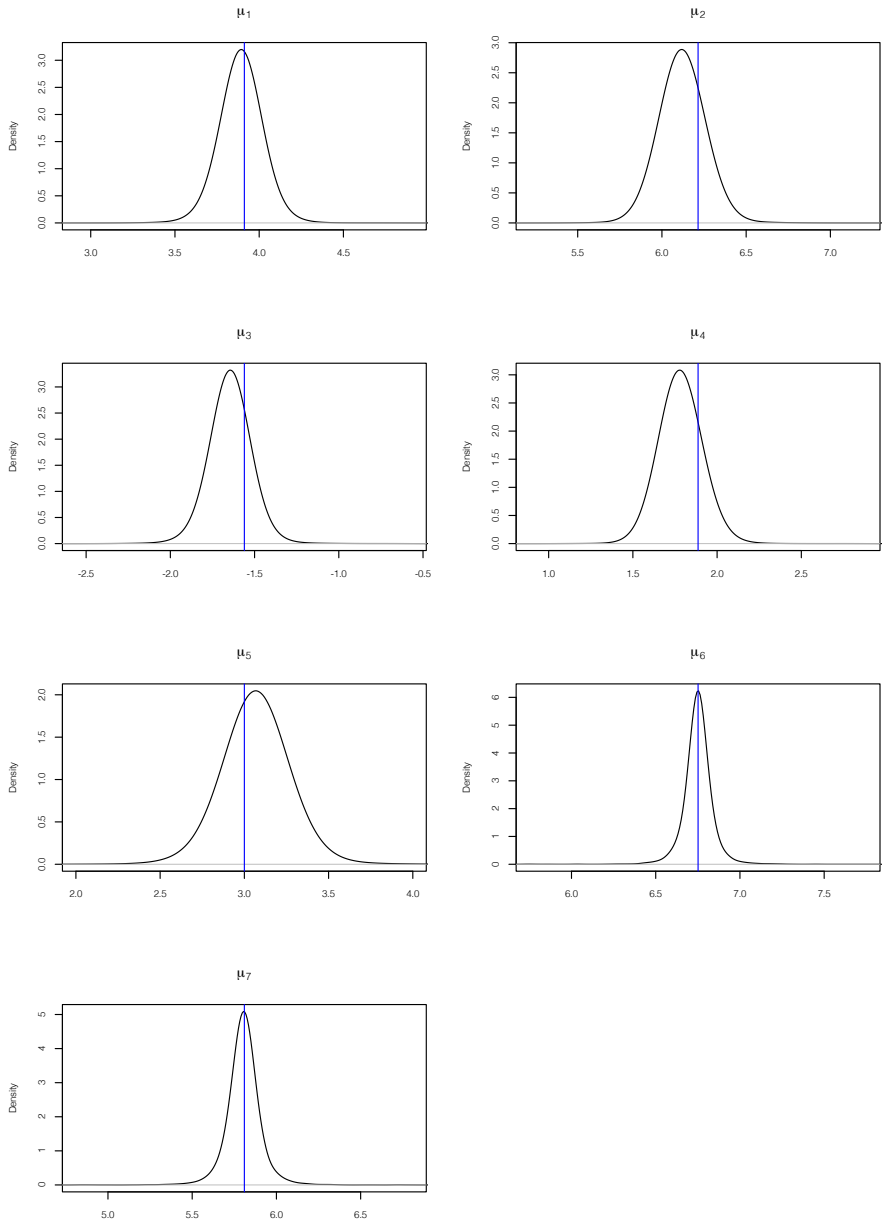


**Figure B.77:** Posterior densities for the hypermean parameters from the second simulation study. The vertical line denotes the real value used in the simulation. Images, from left to right and top to bottom, refer in the order to the hypermeans of $\tilde{\alpha}_0^{(k)}, \tilde{\alpha}_1^{(k)}, \tilde{k}_{ON}^{(k)}, \tilde{k}_{OFF}^{(k)}, \kappa^{(k)}, \mu_\epsilon^{(k)}$ and $\sigma_\epsilon^{(k)}$. Note that, since we did not simulate the hierarchical parameters $\mu_\epsilon^{(k)}$ and $\sigma_\epsilon^{(k)}$ from a hyperprior distribution, the lines for these two parameters refer to the mean of the respective logarithmic hierarchical parameters over the four replicates, i.e. $\frac{1}{4}\left(log(\mu_\epsilon^{(1)}) + log(\mu_\epsilon^{(2)}) + log(\mu_\epsilon^{(3)}) + log(\mu_\epsilon^{(4)})\right)$ and $\frac{1}{4}\left(log(\sigma_\epsilon^{(1)}) + log(\sigma_\epsilon^{(2)}) + log(\sigma_\epsilon^{(3)}) + log(\sigma_\epsilon^{(4)})\right)$, respectively.

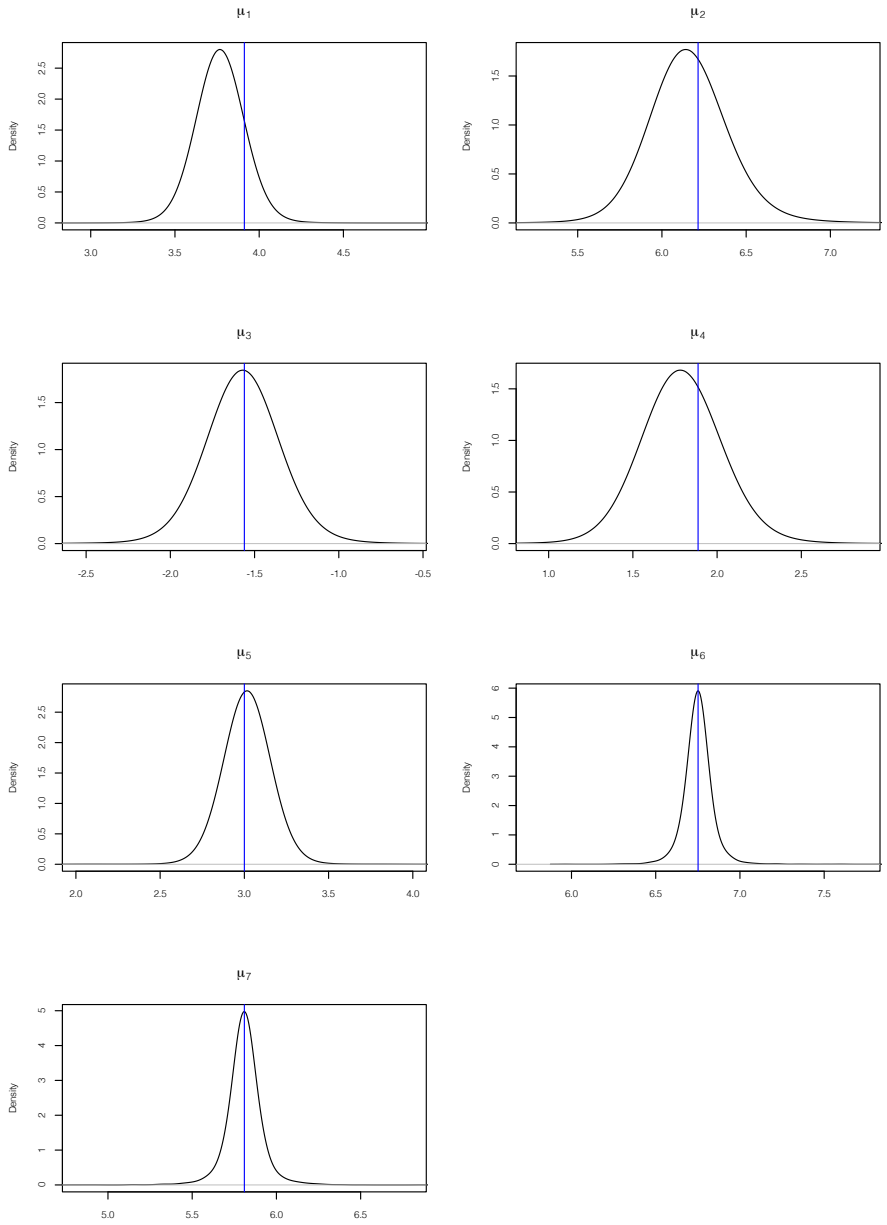


**Figure B.78:** Posterior densities for the hypermean parameters from the third simulation study. The vertical line denotes the real value used in the simulation. Images, from left to right and top to bottom, refer in the order to the hypermeans of $\tilde{\alpha}_0^{(k)}, \tilde{\alpha}_1^{(k)}, \tilde{k}_{ON}^{(k)}, \tilde{k}_{OFF}^{(k)}, \kappa^{(k)}, \mu_\epsilon^{(k)}$ and $\sigma_\epsilon^{(k)}$. Note that, since we did not simulate the hierarchical parameters $\mu_\epsilon^{(k)}$ and $\sigma_\epsilon^{(k)}$ from a hyperprior distribution, the lines for these two parameters refer to the mean of the respective logarithmic hierarchical parameters over the four replicates, i.e. $\frac{1}{4}\left(log(\mu_\epsilon^{(1)}) + log(\mu_\epsilon^{(2)}) + log(\mu_\epsilon^{(3)}) + log(\mu_\epsilon^{(4)})\right)$ and $\frac{1}{4}\left(log(\sigma_\epsilon^{(1)}) + log(\sigma_\epsilon^{(2)}) + log(\sigma_\epsilon^{(3)}) + log(\sigma_\epsilon^{(4)})\right)$, respectively.

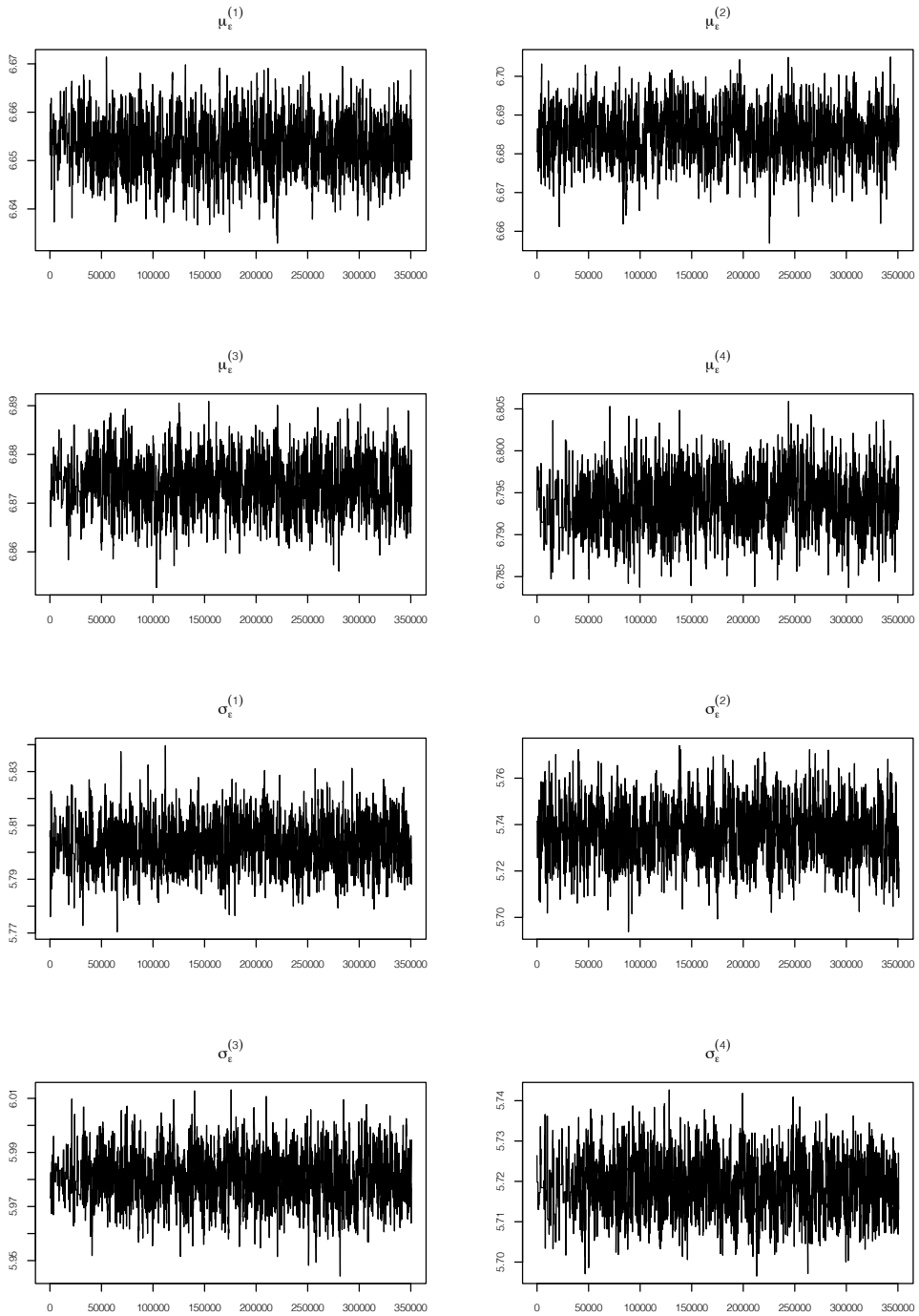


**Figure B.79:** Trace plots, in the log space, for the posterior chains of mean and standard deviation of the measurement error in the four replicates, $log(\mu_\epsilon^{(k)})$ and $log(\sigma_\epsilon^{(k)})$, $k = 1, ..., 4$. Parameters are inferred, via a Bayesian hierarchical analysis, from the background data, as described in the measurement error exploratory analysis in Section 8.1.

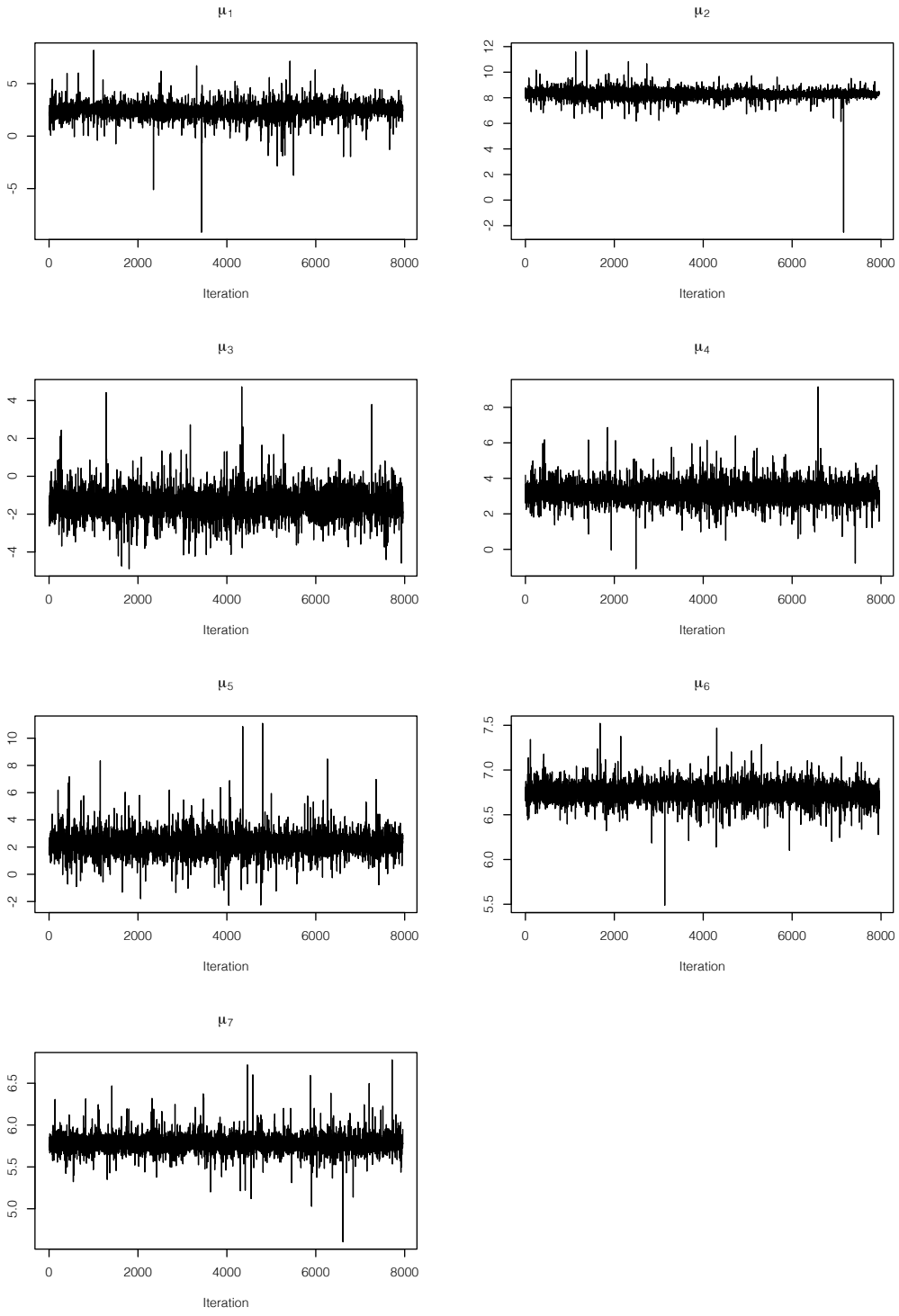


**Figure B.80:** Trace plots for the posterior chains of the hypermean parameters, for cells under no induction. Chains are thinned with tinning factor equals to 100.

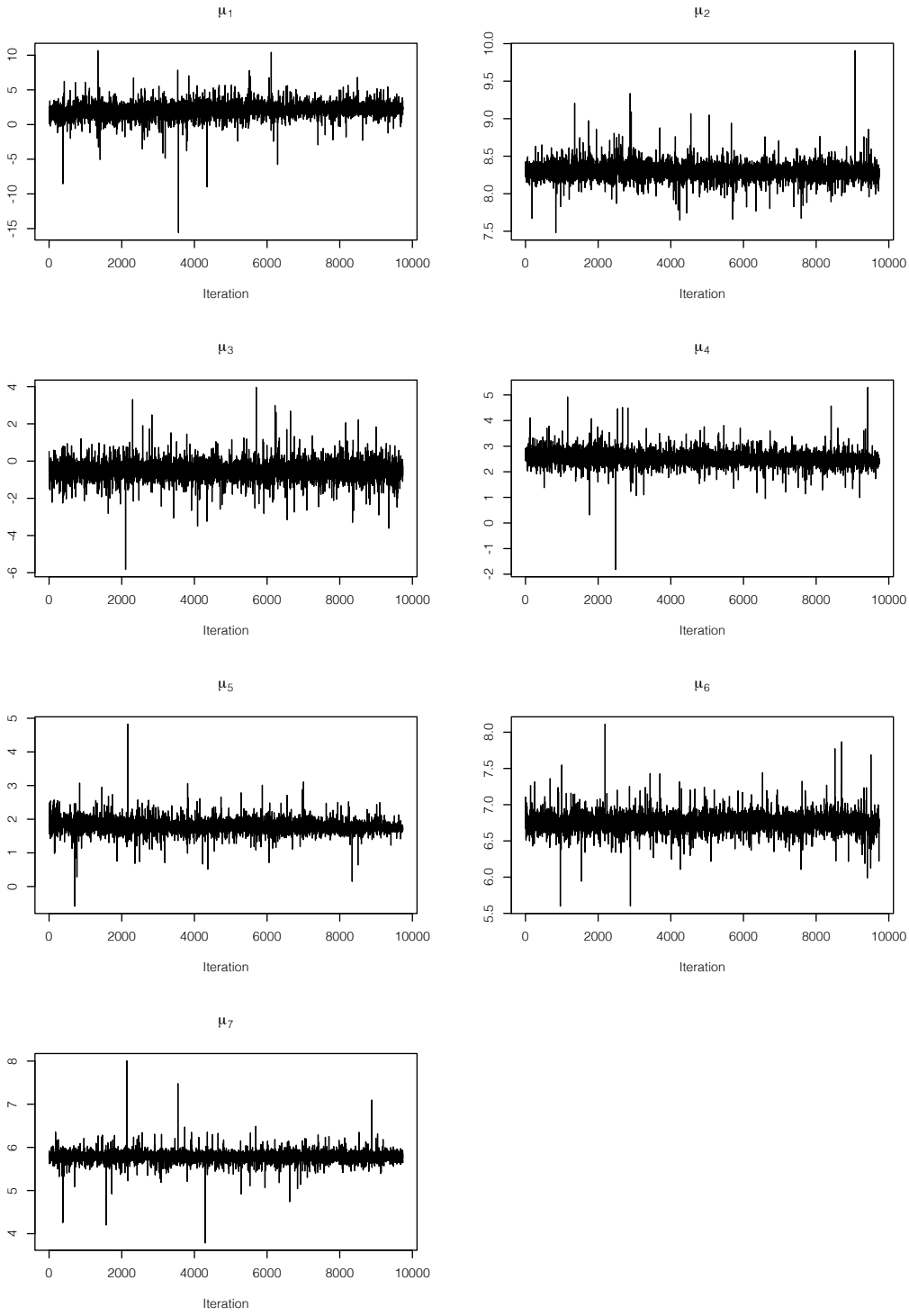


**Figure B.81:** Trace plots for the posterior chains of the hypermean parameters, for cells under the induction by *tetracycline* at 5 ng/ml. Chains are thinned with tinning factor equals to 100.

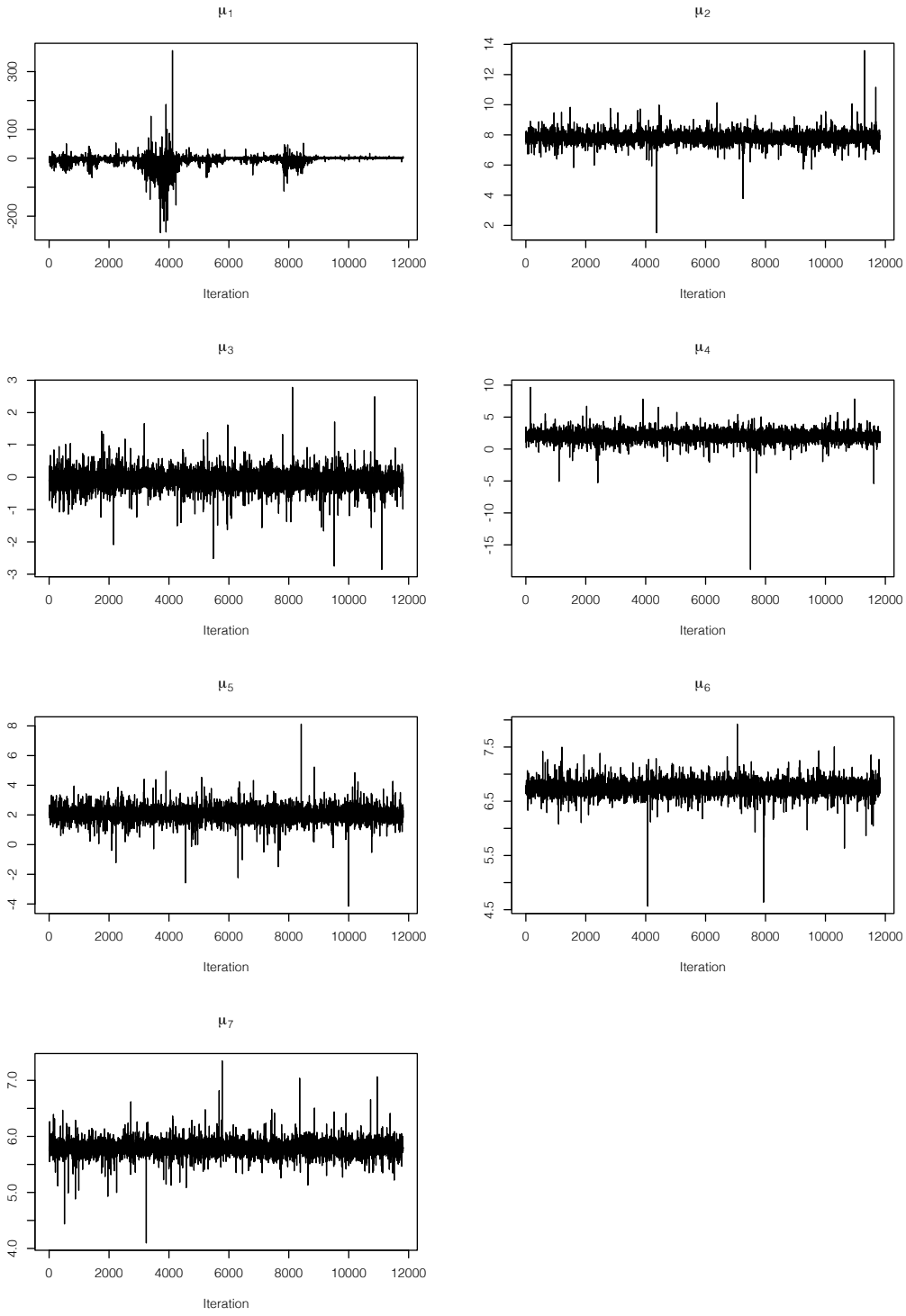


**Figure B.82:** Trace plots for the posterior chains of the hypermean parameters, for cells under the induction by *tetracycline* at 10 ng/ml. Chains are thinned with tinning factor equals to 100.

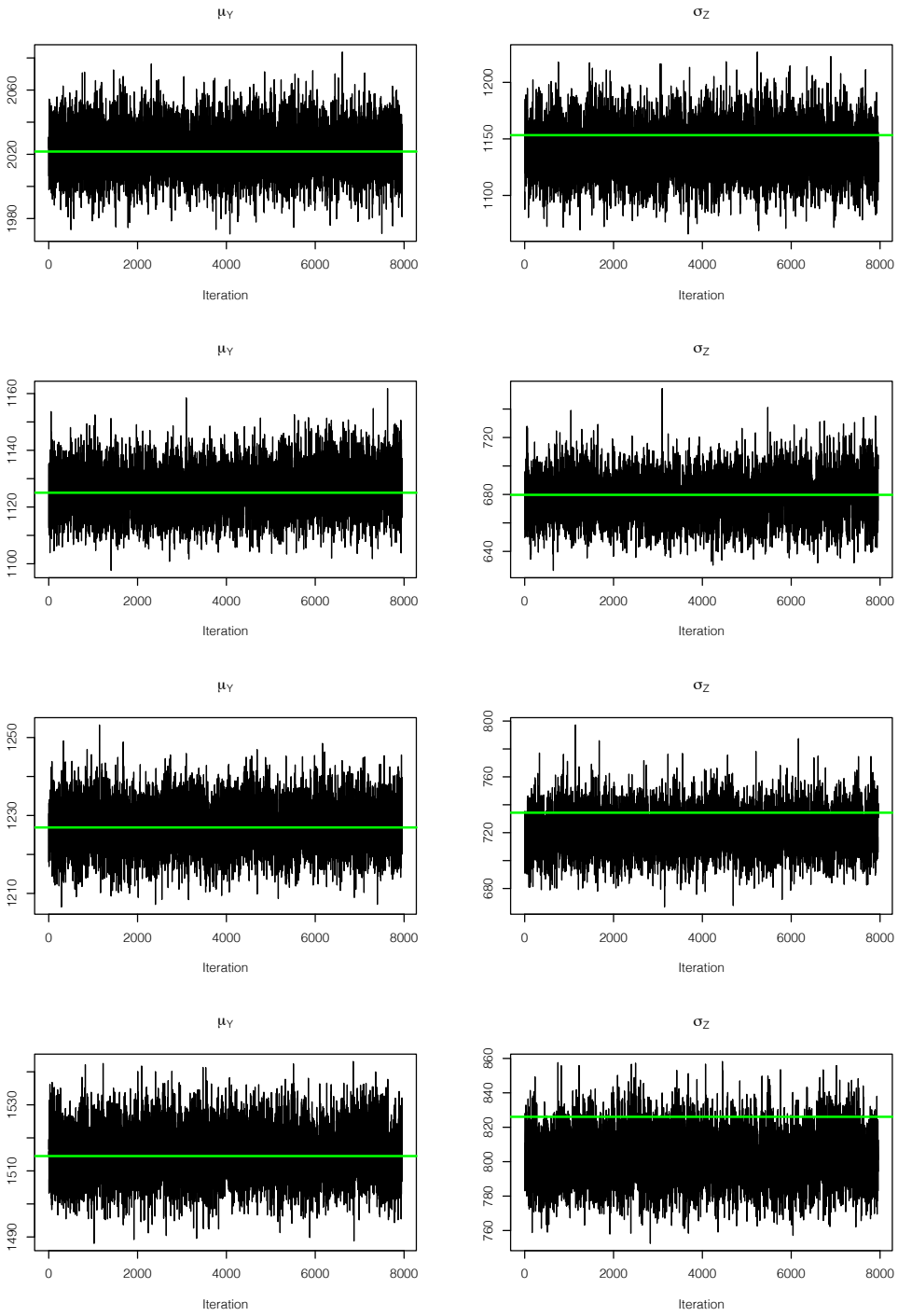


**Figure B.83:** Trace plots for the hierarchical posterior chains of $\mu_Y^{(k)}$ (right column) and $\sigma_Y^{(k)}$ (left column), for cells under no induction. The horizontal green lines denote the mean and standard deviation of the observations in the respective experiment. Rows, from top to bottom, correspond to experiments 1 to 4. Chains are thinned with tinning factor equals to 100.

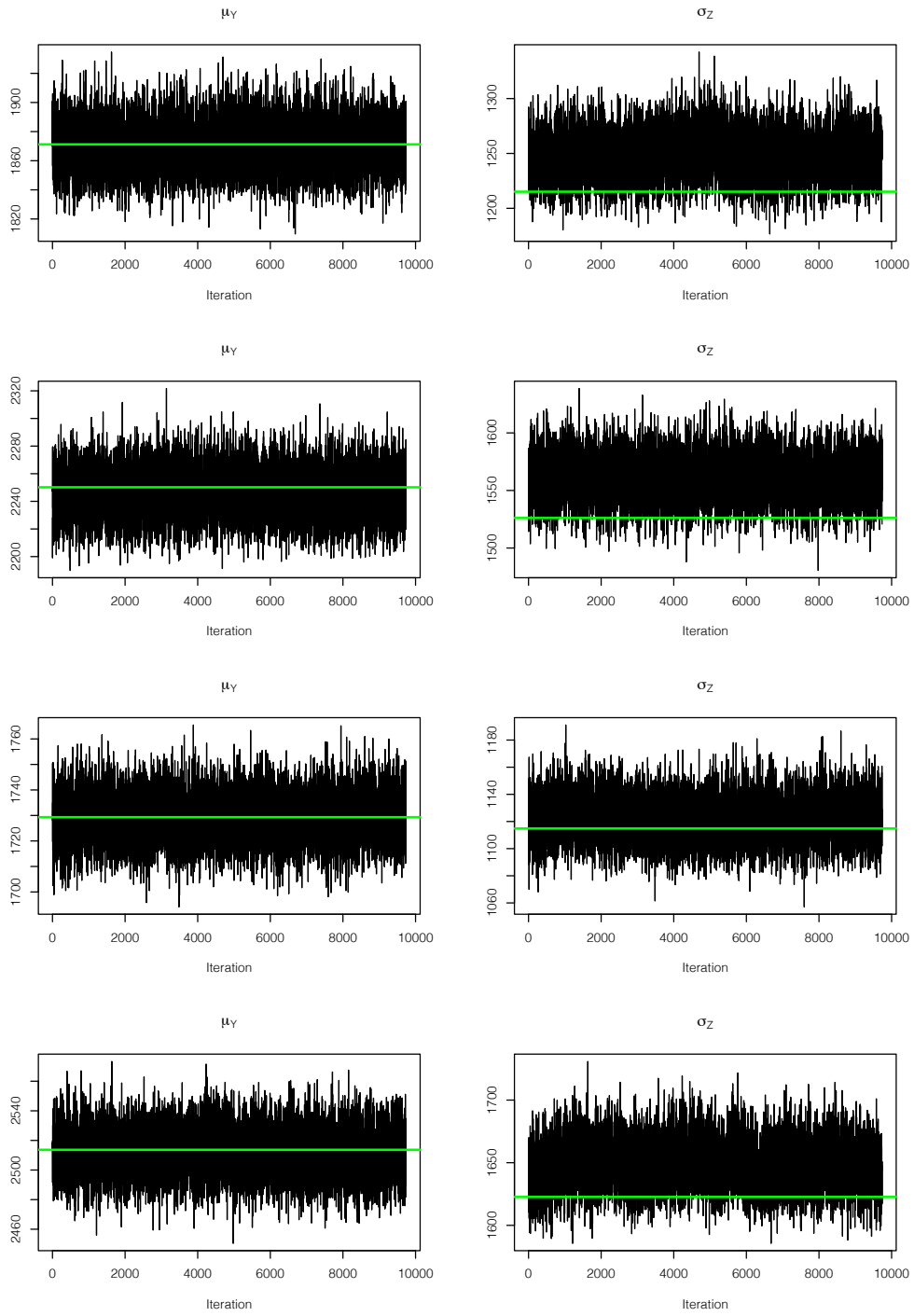


**Figure B.84:** Trace plots for the hierarchical posterior chains of $\mu_Y^{(k)}$ (right column) and $\sigma_Y^{(k)}$ (left column), or cells under the induction by *tetracycline* at 5 ng/ml. The horizontal green lines denote the mean and standard deviation of the observations in the respective experiment. Rows, from top to bottom, correspond to experiments 1 to 4. Chains are thinned with tinning factor equals to 100.

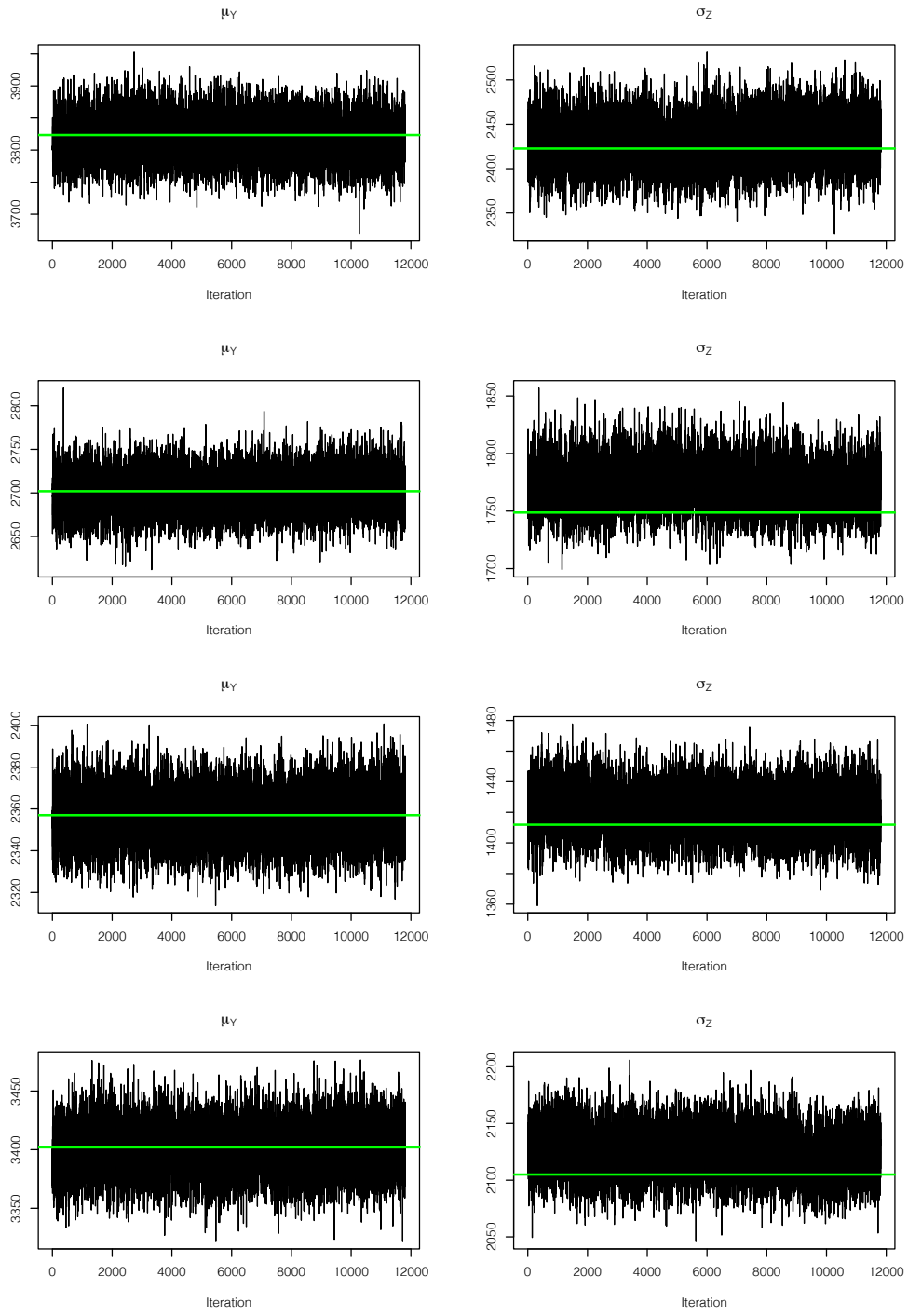


**Figure B.85:** Trace plots for the hierarchical posterior chains of $\mu_Y^{(k)}$ (right column) and $\sigma_Y^{(k)}$ (left column), or cells under the induction by *tetracycline* at 10 ng/ml. The horizontal green lines denote the mean and standard deviation of the observations in the respective experiment. Rows, from top to bottom, correspond to experiments 1 to 4. Chains are thinned with tinning factor equals to 100.

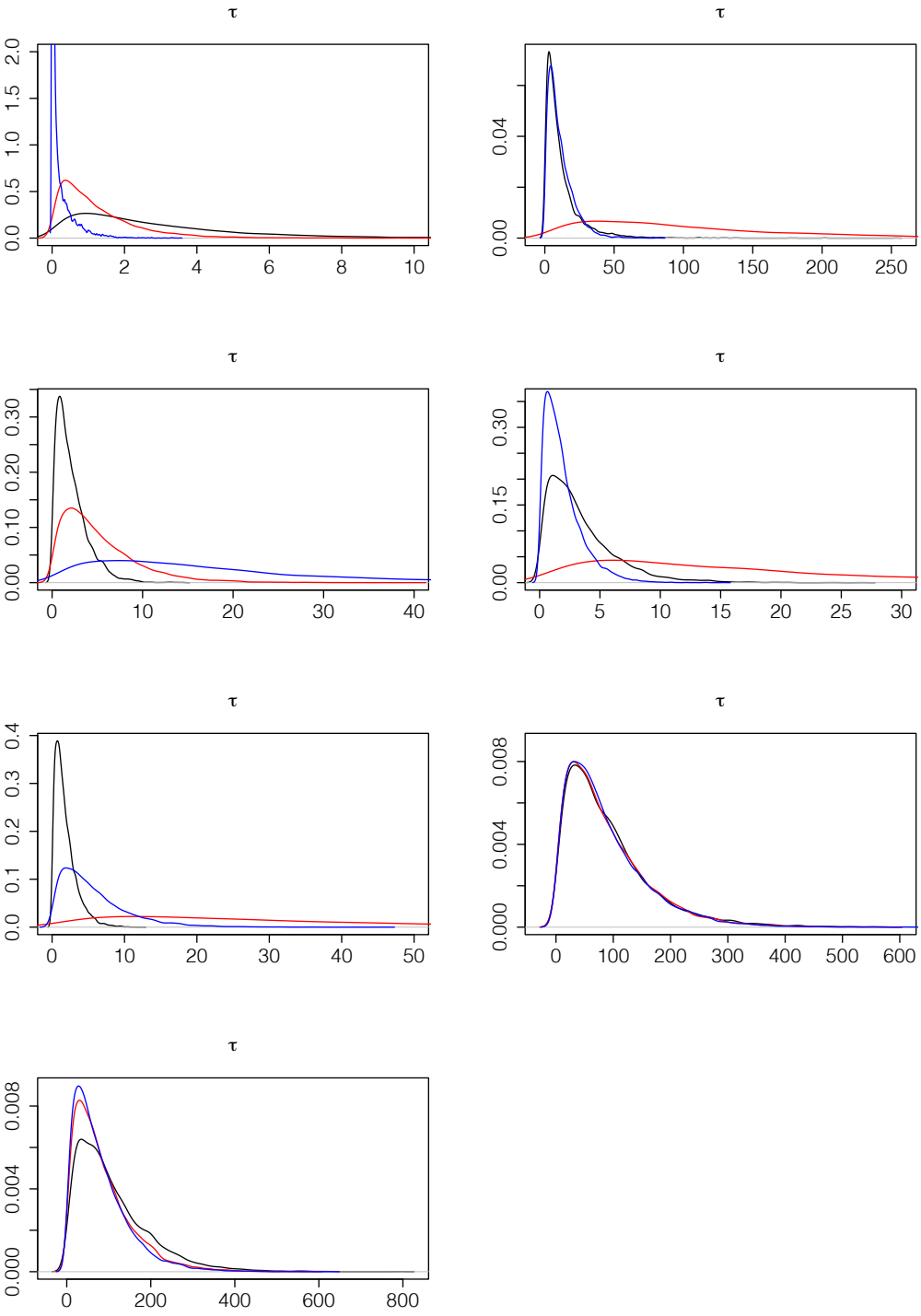


**Figure B.86:** Posterior densities for the hyperprecision parameters in the three experimental conditions: the black, red and blue lines indicate the 0, 5 and 10 ng/ml levels of induction, respectively. Images, from left to right and top to bottom, refer in the order to the hypermeans of $\tilde{\alpha}_0^{(k)}, \tilde{\alpha}_1^{(k)}, \tilde{k}_{ON}^{(k)}, \tilde{k}_{OFF}^{(k)}, \kappa^{(k)}, \mu_\epsilon^{(k)}$ and $\sigma_\epsilon^{(k)}$.

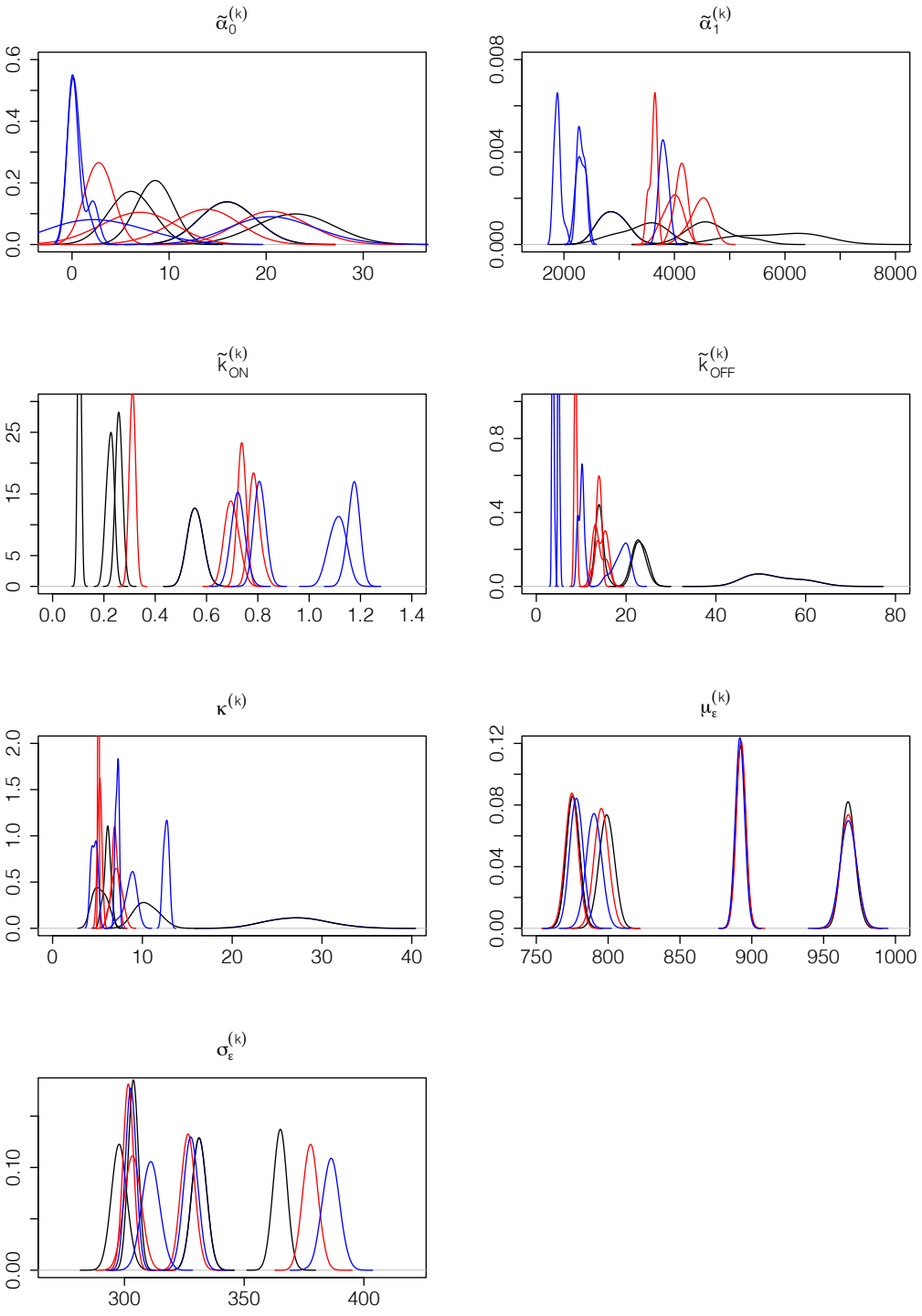


**Figure B.87:** Posterior densities for the hierarchical parameters of the fours experiments in the three experimental conditions: the black, red and blue lines indicate the 0, 5 and 10 ng/ml levels of induction, respectively. Images, from left to right and top to bottom, refer in the order to the hypermeans of $\tilde{\alpha}_0^{(k)}, \tilde{\alpha}_1^{(k)}, \tilde{k}_{ON}^{(k)}, \tilde{k}_{OFF}^{(k)}, \kappa^{(k)}, \mu_\epsilon^{(k)}$ and $\sigma_\epsilon^{(k)}$.

# Bibliography


Balanov, A. G., Janson, N. B., and Schöll, E. (2004). Control of noise-induced oscillations by delayed feedback. *Physica D: Nonlinear Phenomena*, **199**(1), 1–12.

Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics*, **37**, 1554–1563.

Biggin, M. D. (2011). Animal transcription networks as highly connected, quantitative continua. *Developmental Cell*, **21**, 611–626.

Calaway, R., Revolution Analytics, Weston, S., and Tenenbaum, D. (2016). *doParallel: Foreach parallel adaptor for the 'parallel' package. R package version 1.0.10.* `http://CRAN.R-project.org/package=doParallel`.

Claydon, A. J. and Beynon, R. (2012). Proteome dynamics: revisiting turnover with a global perspective. *Molecular & Cellular Proteomics*, **11**, 1551–1565.

Dar, R. D., Razooky, B. S., Singh, A., Trimeloni, T. V., McCollum, J. M., Cox, C. D., Simpson, M. L., and Weinberger, L. S. (2012). Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proceedings of the National Academy of Sciences*, **109**(43), 17454–17459.

Dinkova-Kostova, A. T., Holtzclaw, D. W., Cole, R. N., Itoh, K., Wakabayashi, N., Katoh, Y., Yamamoto, M., and Talalay, P. (2002). Direct evidence that sulfhydryl groups of Keap1 are the sensors regulating induction of phase 2 enzymes that protect against carcinogens and oxidants. *Proc Natl Acad Sci*, **99**, 11908–11913.

Down, J. E. and Riggs, D. S. (1964). A comparison of estimates of Michaelis-Menten kinetic constants from various linear transformations. *The Journal of Biological Chemistry*, **240**, 863–869.

Du, C. and Bretschneider, T. (2013). *Updated CellTracker software manual.* `http://www2.warwick.ac.uk/fac/sci/systemsbiology/staff/bretschneider/celltracker/celltracker-manual.pdf`.

Durham and Gallant (2002). Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes. *Journal of Business & Economic Statistics*, **3**, 335–338.

Elerian, O., Chib, S., and Shephard, N. (2001). Likelihood inference for discretely observed nonlinear diffusions. *Econometrica*, **69**, 959–993.

Finkenstädt, B., Woodcock, D. J., Komorowski, M., Harper, C. V., Davis, J. R. E., White, M. R. H., and Rand, D. A. (2013). Quantifying intrinsic and extrinsic noise in gene transcription using the linear noise approximation: an application to single cell data. *The Annals of Applied Statistics*, **7**, 1960–1982.

Fisher, R. A. (1935). The design of experiments. *Hafner Publishing Company, New York*.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian analysis*, **1**(3), 515–534.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., Bornkamp, B., Maechler, M., and Hothorn, T. (2016). *mvtnorm: multivariate normal and t distributions. R package version 1.0-5.* `https://cran.r-project.org/web/packages/mvtnorm/`.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, **81**(25), 2340–2361.

Gillespie, D. T. (1991). Markov processes: an introduction for physical scientists. *Academic Press, San Diego*.

Gillespie, D. T. (2000). The chemical Langevin equation. *Journal of Chemical Physics*, **113**, 297–306.

Goldbeter, A. (1996). Biochemical oscillations and cellular rhythms: the molecular bases of periodic and chaotic behaviour. *Cambridge University Press, Cambridge*.

Golding, I., Paulsson, J., Zawilski, S. M., and Cox, E. C. (2005). Real-time kinetics of gene activity in individual bacteria. *Cell*, **123**, 1025–1036.

Haario, H., Saksman, E., and Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli*, **7**, 223–242.

Hadfield, J. D. (2016). *MCMCglmm: MCMC generalised linear mixed models*. `https://cran.r-project.org/web/packages/MCMCglmm/`.

Hampsey, M., Badri, N. S., Athar, A., Jean, P. L., and Krishnamurthya, S. (2011). Control of eukaryotic gene expression: gene loops and transcriptional memory. *Advances in Enzyme Regulation*, **51**, 118–125.

Hasegawa, H. (2004). Augmented moment method for stochastic ensembles with delayed couplings. I. Langevin model. *Physical Review E*, **70**(2), 021911.

Hastings, W. K. (1970). Monte carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Hebenstreit, D. (2013). Are gene loops the cause of transcriptional noise? *Trends in Genetics*, **29**, 333–338.

Heron, E. A., Finkenstädt, B., and Rand, D. A. (2007). Bayesian inference for dynamic transcriptional regulation; the Hes1 system as a case study. *Bioinformatics*, **23**, 2596–2603.

Hirsch, M. W., Smale, S., and Devaney, R. L. (2013). Differential equations, dynamical systems, and an introduction to chaos. *Academic Press, Waltham*.

Itoh, K., Wakabayashi, N., Katoh, Y., Ishii, T., O'Connor, T., and Yamamoto, M. (2003). Keap1 regulates both cytoplasmic-nuclear shuttling and degradation of Nrf2 in response to electrophiles. *Genes to Cells*, **8**, 379–391.

Johnson, N. L., Kemp, A. W., and Kotz, S. (2005). Univariate discrete distributions. *John Wiley & Sons, Hoboken*.

Kaern, M., Elston, T. C., Blake, W. J., and Collins, J. J. (2005). Stochasticity in gene expression: from theories to phenotypes. *Nat Rev Genet*, **6**, 451–464.

Kaspar, J. W. and Jaiswal, A. K. (2010). Tyrosine phosphorylation controls nuclear export of Fyn, allowing Nrf2 activation of cytoprotective gene expression. *FASEB J*, **25**, 1076–1087.

Kim, J. K. and Marioni, J. C. (2013). Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. *Genome Biology*, **14**(1), 1–12.

Ko, C. H., Yamada, Y. R., Welsh, D. K., Buhr, E. D., Liu, A. C., Zhang, E. E., Ralph, M. R., Kay, S. A., Forger, D. B., and Takahashi, J. S. (2010). Emergence of noise-induced oscillations in the central circadian pacemaker. *PLoS Biol*, **8**(10), e1000513.

Kobayashi, A., Kang, M., Okawa, H., Ohtsuji, M., Zenke, Y., Chiba, T., Igarashi, K., and Yamamoto, M. (2004). Oxidative stress sensor Keap1 functions as an adaptor for Cul3-based E3 ligase to regulate proteasomal degradation of Nrf2. *Molecular and Cellular Biology*, **24**, 7130–7139.

Lee, J. M., Li, J., Johnson, D. A., Stein, T. D., Kraft, A. D., Calkins, M. J., Jakel, R. J., and Johnson, J. A. (2005). Nrf2, a multi-organ protector? *FASEB J*, **19**, 1061–1066.

Lesaffre, E. and Lawson, A. B. (2012). Bayesian biostatistics. *John Wiley & Sons, Chichester*.

Lo, S. C. and Hannink, M. (2006). PGAM5, a Bcl-XL-interacting protein, is a novel substrate for the redox-regulated Keap1-dependent ubiquitin ligase complex. *J Biol Chem*, **281**, 37893–37903.

Luenberger, D. G. D. G. (1979). Introduction to dynamic systems; theory, models, and applications. *John Wiley & Sons, New York*.

Malhotra, D., Portales-Casamar, E., Singh, A., Srivastava, S., Arenillas, D., Happel, C., Shyr, C., Wakabayashi, N., Kensler, T. W., Wasserman, W. W., and Biswal, S. (2010). Global mapping of binding sites for Nrf2 identifies novel targets in cell survival response through ChIP-Seq profiling and network analysis. *Nucl Acids Res*, **38**, 5718–5734.

Mapendano, C. K., Lykke-Andersen, S., Kjems, J., Bertrand, E., and Jensen, T. H. (2010). Crosstalk between mRNA 3' end processing and transcription initiation. *Molecular Cell*, **40**, 410–422.

Metropolis, N. and Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, **44**, 335–341.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, **21**, 1087–1092.

Mills, T. C. (1991). Time series techniques for economists. *Cambridge University Press, Cambridge*.

Momiji, H. and Monk, N. A. M. (2009). Oscillatory Notch-pathway activity in a delay model of neuronal differentiation. *Physical Review E*, **80**, 021930.

Muller, K. E. (2001). Computing the confluent hypergeometric function, M (a, b, x). *Numerische Mathematik*, **90**(1), 179–196.

Munsky, B., Neuert, G., and van Oudenaarden, A. (2012). Using gene expression noise to understand gene regulation. *Science*, **336**(6078), 183–187.

Pace, L. and Salvan, A. (1997). Principles of statistical inference: from a neo-Fisherian perspective. *World scientific, Singapore*.

Paulsson, J. (2005). Models of stochastic gene expression. *Physics of life reviews*, **2**(2), 157–175.

Peccoud, J. and Ycart, B. (1995). Markovian modeling of gene-product synthesis. *Theoretical population biology*, **48**(2), 222–234.

Pi, J., Bai, Y., Reece, J. M., Williams, J., Liu, D., Freeman, M. L., Fahl, W. E., Shugar, D., Liu, J., Qu, W., Collins, S., and Waalkes, M. P. (2007). Molecular mechanism of human Nrf2 activation and degradation: role of sequential phosphorylation by protein kinase CK2. *Free Radical Biol Med*, **42**, 1797–1806.

Plummer, M., Best, N., Cowles, K., Vines, K., Sarkar, D., Bates, D., Almond, R., and Magnusson, A. (2016). *coda: output analysis and diagnostics for MCMC. R package version 0.18-1.* `https://cran.r-project.org/web/packages/coda/`.

R Core Team (2016). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. `https://www.R-project.org`.

Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y., and Tyagi, S. (2006). Stochastic mRNA synthesis in mammalian cells. *PLoS Biology*, **4**(10), 1–13.

Rajala, T., Häkkinen, A., Healy, S., Yli-Harja, O., and Ribeiro, A. S. (2010). Effects of transcriptional pausing on gene expression dynamics. *PLoS Comput Biol*, **6**(3), 1–12.

Roberts, G. O. and Rosenthal, J. S. (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, **18**(2), 349–367.

Schleif, R. (1992). DNA looping. *Annual review of biochemistry*, **61**(1), 199–223.

Shahrezaei, V. and Swain, P. S. (2008). Analytical distributions for stochastic gene expression. *Proceedings of the National Academy of Sciences*, **105**(45), 17256–17261.

Singh, A., Vargas, C. A., and Karmakar, R. (2013). Stochastic analysis and inference of a two-state genetic promoter model. *American Control Conference*, pages 4563–4568.

So, L.-h., Ghosh, A., Zong, C., Sepulveda, L. A., Segev, R., and Golding, I. (2011). General properties of transcriptional time series in Escherichia coli. *Nature Genetics*, **43**, 554–560.

Soetaert, K., Petzoldt, T., Setzer, R. W., and odepack authors (2016). *deSolve: solvers for initial value problems of differential equations (ODE, DAE, DDE). R package version 1.14.* `https://cran.r-project.org/web/packages/deSolve/`.

Stathopoulos, V. and Girolami, M. A. (2013). Markov chain Monte Carlo inference for Markov jump processes via the linear noise approximation. *Phil Trans R Soc A*, **371**, 20110541.

Steuer, R. (2004). Effects of stochasticity in models of the cell cycle: from quantized cycle times to noise-induced oscillations. *Journal of theoretical biology*, **228**(3), 293–301.

Strogatz, S. H. (2014). Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering. *Westview press, Boulder*.

Suter, D. M., Molina, N., Gatfield, D., Schneider, K., Schibler, U., and Naef, F. (2011). Mammalian genes are transcribed with widely different bursting kinetics. *Science*, **332**(6028), 472–474.

Taguchi, K. and M, H. M. M. Y. (2011). Molecular mechanisms of the Keap1- Nrf2 pathway in stress response and cancer evolution. *Genes to Cells*, **16**, 123–140.

Takeda, K., Komuro, Y., Hayakawa, T., Oguchi, H., Ishida, Y., Murakami, S., Noguchi, T., Kinoshita, H., Sekine, Y., Iemura, S., Natsume, T., and H, H. I. (2009). Mitochondrial phosphoglycerate mutase 5 uses alternate catalytic activity as a protein serine/threonine phosphatase to activate ASK1. *Proc Natl Acad Sci USA*, **106**, 12301–12305.

Tanigawa, S., Fujii, M., and Hou, D. X. (2007). Action of Nrf2 and Keap1 in ARE-mediated NQO1 expression by quercetin. *Free Radical Biol Med*, **42**, 1690–1703.

Tian, T., Burrage, K., Burrage, P. M., and Carletti, M. (2007). Stochastic delay differential equations for genetic regulatory networks. *Journal of Computational and Applied Mathematics*, **205**(2), 696–707.

Wilkinson, D. J. (2012). Stochastic modelling for systems biology, second edition. *Chapman & Hall/CRC Mathematical and Computational Biology Series, Boca Raton*.

Wills, Q. F., Livak, K. J., Tipping, A. J., Enver, T., Goldson, A. J., Sexton, D. W., and Holmes, C. (2013). Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nat Biotech*, **31**, 748–752.

Xue, M., Momiji, H., Rabbani, N., Bretschneider, T., Rand, D. A., and Thornalley, P. J. (2015a). Frequency modulated translocational oscillations of Nrf2, a transcription factor functioning like a wireless sensor. *Biochemical Society Transactions*, **43**, 669–673.

Xue, M., Momiji, H., Rabbani, N., Barker, G., Bretschneider, T., Shmygol, A., Rand, D. A., and Thornalley, P. J. (2015b). Frequency modulated translocational oscillations of Nrf2 mediate the antioxidant response element cytoprotective transcriptional response. *Antioxidants & Redox Signaling*, **23**, 613–629.

Zopf, C. J., Quinn, K., Zeidman, J., and Maheshri, N. (2013). Cell-cycle dependence of transcription dominates noise in gene expression. *PLoS Comput Biol*, **9**(7), 1–12.